Local asymptotics of cross-validation around the optimal model

Guillaume Maillard

Local asymptotics of cross-validation around the optimal model

Guillaume Maillard
guillaume.maillard@ensai.fr

February 10, 2026

Abstract

When used to estimate the risk of a single predictor, the behaviour of cross-validation can often be understood through a central limit theorem. In model selection however, cross-validation is applied simultaneously to many different estimators in order to compare them. Thus, analyzing CV in this context requires a multi-dimensional or functional CLT. Since the mean and variance may vary widely over the model collection, careful attention must be paid to how the process is centered and scaled. In this article, we conduct the first such analysis of cross-validation in the context of least-squares density estimation by Fourier polynomials. Our results characterize the fluctuations of some CV criteria in the vicinity of the optimal model, at the critical scale at which they become significant. Asymptotically, CV is approximately the sum of a deterministic function (the expected risk) and a symmetrized, time-changed Wiener process. For a slowly increasing number of folds V, the variance decreases proportionally to $\frac{1}{V}$: the folds are asymptotically independent. This analysis presents some unusual challenges which we overcome through a combination of tools including strong approximation, concentration inequalities and coupling of Gaussian vectors.

1 Introduction

Cross-validation is a widely-used class of methods for risk estimation and model selection. Despite its popularity, its behaviour is still not completely understood and much practical advice is based on heuristic arguments or numerical simulations. This sometimes leads to disagreement between experts [1], such as on the question of whether V-fold [13, Chapter 5] or leave-1-out [6] has better performance.

The fluctuations of CV along a given sequence of predictors (indexed by the sample size) can be assessed by computing the variance and establishing a central limit theorem. Precise variance computations were carried out in various settings in [6, 7, 2] among others - see the review paper [1] for a more complete overview. On the other hand, central limit theorems were proved in a general setting in [10, 3] and most recently in [4]. Together, these results can provide accurate information about the deviations of cross-validation when estimating the risk of a single predictor, or sequence of predictors.

However, in many cases, cross-validation is applied simultaneously to a whole class of estimators rather than just one: for example, empirical risk minimizers on a nested sequence of models, or Lasso predictors with various values of the penalty parameter. This class of predictors may change with the sample size: for example, models will typically be allowed to grow bigger and penalty parameters to become smaller as n grows. Practically relevant quantities, such as the model or parameter selected by cross-validation and its risk, depend on the whole joint distribution of cross-validation applied to the given class of predictors. The behaviour of cross-validation in the neighbourhood of the optimal model is particularly significant as

we typically want and expect model selection estimators to concentrate around the optimal model, at least in the sense of having nearly optimal risk. For cross-validation based model selection, such results, known as oracle inequalities, have been established in regression [20, 18] and least-squares density estimation [2]. However, oracle inequalities do not tell the whole story as they "only" provide an upper bound on the rate of concentration of the selected estimator. Ideally, we would like to characterize the fluctuations of the CV selected model and estimator around their optimal ("oracle") counterparts, as central limit theorems do for the CV risk estimator at a single model. This raises the question of establishing a multi-dimensional or functional CLT for cross-validation, at least locally around the optimal model/parameter.

In this article, we focus on the special case of least-squares density estimation, with a collection of linear models generated by cosine functions with increasing frequencies, a model collection which is both practically useful and allows for explicit computations. The study of cross-validation in that context has several original characteristics which distinguish it from textbook applications of empirical process theory, even in the simplest case of simple validation. The hold-out risk estimator, centered at the optimal model, can be written as

$$\mathrm{HO}_T(k) - \mathrm{HO}_T(k*) = \left\|\hat{s}_k^T - s\right\|^2 - \left\|\hat{s}_{k*}^T - s\right\|^2 - 2\left(P_n^{T^c} - P\right)(\hat{s}_k^T - \hat{s}_{k*}^T),$$

where $\hat{s}_k^T$ is the empirical risk minimizer on the $k-$th model, computed using the training data, s is the true density with corresponding probability measure P and $P_n^{T^c}$ is the empirical measure on the test data. Hence, it can be remarked that, though $\mathrm{HO}_T(k) - \mathrm{HO}_T(k*)$ is an empirical process *conditionally on the training data*,

- The class of functions, of the form $(\hat{s}_k)_{k\in K_n}$, is random, depends on n and the estimators $\hat{s}_k$ range over models of unbounded dimension.
- The variance, $\mathrm{Var}_P\left(\hat{s}_k^T - \hat{s}_{k*}^T\right)$, converges to 0 since, for k close enough to the optimum, $\hat{s}_k$ and $\hat{s}_{k_*}$ converge to s.
- The process, suitably renormalized, may fail to converge in distribution. If so, we want to approximate it by a sequence of "simple" Gaussian process.

In order to find a Gaussian approximation in this non-standard setting, we use the celebrated Komlós-Major-Tusnády (KMT) Theorem [15], which provides a coupling between the empirical process $(\sqrt{n}(F_n - F)(t))_{t\in\mathbb{R}}$, where F is the cdf and F_n the empirical cdf, and a Brownian bridge process $B_n(t)$, such that

$$\mathbb{E}\left[\left\|\sqrt{n}(F_n - F) - B_n \circ F\right\|_\infty\right] \leq C\frac{\log n}{\sqrt{n}},$$

for some numerical constant C. The KMT theorem can be used to approximate a general empirical process of the form $\sqrt{n}(P_n - P)(f)_{f\in\mathcal{F}}$ by the Gaussian process

$$\left(\int B_n(F(t))df(t)\right)_{f\in\mathcal{F}}, \tag{1}$$

whenever the functions in $\mathcal{F}$ are of bounded variation. This construction was first used in [12] to prove a law of the iterated logarithm for kernel density estimators, then in a more general empirical process setting in [19]. Other methods were later used to obtain strong approximations to general empirical processes under different assumptions [14, 5, 8].

However, the simple construction (1) is sufficient in the present setting. Its main technical advantage, in addition to its simplicity (modulo the KMT construction, used as a black box), is that the underlying Brownian bridge process can clearly be constructed independently of the class of estimators $\hat{s}_k^T$.

The resulting Gaussian process has a complicated, random covariance function. By approximating it through explicit computations and concentration inequalities, we show that the Gaussian process itself can be approximated by a symmetrized Brownian motion changed in time, W_{g_n}, where g_n is an increasing function. In combination with concentration inequalities for the risk, $\left\|\hat{s}_k^T - s\right\|^2$, this implies that the hold-out process, centered at the optimal model and appropriately rescaled, is approximately the sum of the non-negative function f_n and the zero-mean Gaussian process W_{g_n}. The same result applies to "incomplete" V-fold cross-validation (Definition 3), with g_n replaced by g_n/V. Both functions are deterministic: as a consequence, the distribution of the hold-out is approximately independent from the training data, and the folds of cross-validation are approximately independent. We establish upper and lower bounds on the functions f_n, g_n which guarantee that f_n, W_{g_n} do not vanish as $n \to +\infty$ and remain bounded /tight, at least when restricted to intervals of non-vanishing length containing the optimal parameter. At larger scales and outside these intervals, the limiting behaviour is trivial in the sense that the (centered and rescaled) hold-out concentrates around a deterministic function. Thus, our results characterize the critical scale at which randomness appears in the asymptotic.

These results hold under the assumption that the sequence of Fourier coefficients of the true density, θ_j, decays at some polynomial rate $j^{-\alpha}$ ($\alpha > \frac{3}{2}$) and is not too irregular in the sense that the lower bound $|\theta_j| \geq cj^{-\alpha}$ holds for some constant c on a non-negligible subset of the integers.

2 Setting

In this article, simple validation and cross-validation are studied in the context of least-squares density estimation. Here, we present this context in a mathematically precise manner.

2.1 L^2 density estimation using empirical orthogonal series

Let $s \in L^2([0;1])$ be a probability density function. Given a sample $X_1, \ldots, X_n$ drawn according to the density s, the L^2 density estimation problem consists in constructing an estimator $\hat{s}_n$ that approaches s in terms of the L^2 norm. Although it is not obvious at first glance (this is not true for the other L^p norms), this non-parametric density estimation problem can be reformulated as a risk minimization problem, with a contrast function: $\gamma(t,x) = \|t\|^2 - 2t(x)$, which yields the *risk* $\mathbb{E}[\gamma(t,X)] = \|t\|^2 - 2\int s(x)t(x)dx = \|t-s\|^2 - \|s\|^2$. It follows that s is indeed the minimizer of the risk corresponding to the γ contrast function, and furthermore the *excess risk* $\ell(s,t) := \mathbb{E}[\gamma(t,X)] - \mathbb{E}[\gamma(s,X)]$ coincides with the squared L^2 norm:

$$\ell(s,t) = \|t-s\|^2 .$$

As a result, it is possible to construct an *empirical risk estimator*,

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,X_i) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n}t(X_i),$$

which can in particular be used to perform cross-validation.

Here we will consider a special class of non-parametric estimators, the *empirical orthogonal series* estimators [11, Chapter 7] on a trigonometric basis. To ease the presentation, we consider only cosine functions, which is equivalent to assuming that s is symmetrical with respect to $\frac{1}{2}$. This restriction is of no fundamental importance — it is reasonable to conjecture that the results remain valid with the complete trigonometric basis.

For every $j \in \mathbb{N}^*$, let $\psi_j : x \mapsto \sqrt{2}\cos(2\pi jx)$ and let $\psi_0 : x \mapsto 1$. The collection $(\psi_j)_{j\in\mathbb{N}}$ is an orthonormal basis of the subset of $L^2([0;1])$ of functions symmetrical with respect to $\frac{1}{2}$.

Let $D_n = (X_1, ..., X_n)$ be a sample. For any $n \in \mathbb{N}$ and any $T \subset \{1, \dots, n\}$, we will denote, for any real-valued measurable function t,

$$P_n^T(t) = \frac{1}{|T|}\sum_{i\in T} t(X_i).$$

Consider the estimators defined as follows.

Definition 1. *For all $k \in \mathbb{N}$ and all $T \subset \{1, \dots, n\}$,*

$$\hat{s}_k^T = \sum_{j=0}^{k} P_n^T(\psi_j)\psi_j,$$

where $\psi_0 = 1$ and for all $j \geq 1$, $\psi_j(x) = \sqrt{2}\cos(2\pi jx)$.

The estimators $\hat{s}_k^T$ are empirical risk minimizers on the *models*

$$E_k = \left\{ \sum_{j=0}^{k} v_j\psi_j : v \in \mathbb{R}^{k+1} \right\}.$$

The problem of parameter choice k is therefore a problem of *model selection* within the model collection $(E_k)_{k\geq 0}$. Here, the models are nested, meaning $E_k \subset E_{k'}$ for every $k \leq k'$.

2.2 Main notation

We summarize the main notation associated with the L^2 density estimation setting, which will be used throughout the article.

- $\mathbb{N}, \mathbb{N}^*$: respectively the set of non-negative integers and the set of positive integers
- $\|\cdot\|$: the norm of the space $L^2([0;1])$
- P, s: the distribution of the observations and its probability density function
- D_n an $n-$sample from P, D_n^T the sub-sample with indices in $T \subset \{1, \dots, n\}$
- P_n, P_n^T the empirical measures associated respectively with the samples D_n, D_n^T.
- ψ_j: the $j-$th function in the cosine basis.
- $\theta_j = \langle s, \psi_j \rangle$: the $j-$th Fourier coefficient of s.
- $\hat{\theta}_j^T = P_n^T(\psi_j)$: estimator of θ_j based on D_n^T.
- $\hat{s}_k^T = 1 + \sum_{j=1}^{k} \hat{\theta}_j^T\psi_j$, the empirical orthogonal series estimator of s based on the k first frequencies.

2.3 Risk estimation for the hold-out

The larger k is, the better the approximation of s by the functions of E_k, but the more difficult it is to estimate the best approximation to s within E_k. The choice of k is therefore subject to a bias-variance trade-off which, if properly carried out, allows adaptation to the smoothness of s, simultaneously reaching the minimax risk on Lipschitz spaces of periodic functions [11, Chapter 7].

Since the risk, except for a constant, is expressed as the expectation of a contrast function

$$P\gamma(\hat{s}_k^T) := E_X\left[\gamma(\hat{s}_k^T, X)\right] = \left\|\hat{s}_k^T - s\right\|^2 - \|s\|^2,$$

it can be estimated by hold-out and cross-validation as in regression and classification. This is the subject of the following definition.

Definition 2. *Let D_n be an i.i.d sample drawn from the distribution $s(x)dx$. Let $n_t \in \{1, \dots, n-1\}$. Let $T \subset \{1, \dots, n\}$ be a subset with cardinality $|T| = n_t$. Then, for all $k \in \mathbb{N}$, we define the hold-out estimator of the risk of $\hat{s}_k$ with training sample indices T by*

$$HO_T(k) = \left\|\hat{s}_k^T\right\|^2 - 2P_n^{T^c}(\hat{s}_k^T).$$

$\mathrm{HO}_T(\cdot)$ is indeed an estimator since the norm $\|\cdot\|$ is computed with respect to a known dominating measure (in this case the Lebesgue measure) and so does not depend on the distribution of X. Moreover,

$$\mathrm{HO}_T(k) = \left\|\hat{s}_k^T - s\right\|^2 - \|s\|^2 - 2(P_n^{T^c} - P)(\hat{s}_k^T) : \tag{2}$$

the hold-out risk estimator can be expressed as the sum of the risk and a centered empirical process. The risk which the hold-out naturally estimates is that of the estimator $\hat{s}_k^T$, which is trained on a sub-sample. If n_t is close enough to n, we expect its risk to be close to that of the same estimator trained on the full sample and this can indeed be proved in our setting (see [2, Lemma 14]).

The hold-out risk estimator depends on the choice of a subset T of $\{1, \dots, n\}$, but its distribution depends only on the cardinality of that subset. The precise choice of a subset T of cardinality n_t will thus play no role in the sequel. We will therefore denote by T any subset of $\{1, \dots, n\}$ of cardinality n_t.

Since the distribution of $\mathrm{HO}_T(\cdot)$ only depends on T through its cardinality n_t, it is possible to construct an estimator with smaller variance by averaging several $\mathrm{HO}_{T_i}(\cdot)$. This is the idea behind cross-validation. In the V-fold scheme presented below, the T_i are chosen such that the test sets T_i^c are disjoint.

Definition 3. *Let n_t, V be integers such that $\frac{V-1}{V} n \leq n_t \leq n-1$. Let $(I_i)_{1 \leq i \leq V}$ be a collection of disjoint subsets of $\{1, \dots, n\}$ of equal cardinality $|I_i| = n - n_t$, chosen independently from the data. For all $i \in \{1, \dots, V\}$, let $T_i = \{1, \dots, n\} \backslash I_i$. Let $\mathcal{T} = (T_1, \dots, T_V)$. The "incomplete" V-fold CV risk estimator is*

$$\mathrm{CV}_{\mathcal{T}}(k) = \frac{1}{V} \sum_{i=1}^{V} HO_{T_i}(k).$$

Similarly to the hold-out, the distribution of $\mathrm{CV}_{\mathcal{T}}(\cdot)$ only depends on n_t, V and in the rest of this article, we will denote by $\mathcal{T}$ any collection $T_1, \dots, T_V$ which satisfies the assumptions of Definition 3. Compared to standard $V-$fold, the cross-validation scheme defined above retains the constraint that the I_i be disjoint and of equal size, but decouples the size of the test sets $|I_i| = n - n_t$ from the number of splits V. In particular, the hold-out (Definition 2) is a special case of Definition 3 (for $V = 1$). As the collection $(I_i)_{1 \leq i \leq V}$ may be "completed" into a partition by adding sets I_j, we shall call $\mathrm{CV}_{\mathcal{T}}(k)$ "incomplete $V-$fold cross-validation".

To analyze the asymptotics of these risk estimators, fix a sequence of integers $(n_t(n))_{n\in\mathbb{N}}$ such that, for all $n \in \mathbb{N}$, $\frac{n}{2} \leq n_t(n) \leq n$, and define $n_v(n) = n - n_t(n)$. In the following, we shall denote $n_t = n_t(n)$ and $n_v = n_v(n)$ for a generic value of n. Whenever n, n_v, n_t appear in the same expression, it will be understood that $n_t = n_t(n)$ and $n_v = n_v(n) = n - n_t(n)$. Since $\mathrm{CV}_{\mathcal{T}}(\cdot)$ can be expressed as an average of $\mathrm{HO}_{T_i}(\cdot)$, we first focus on analyzing the hold-out risk estimator $\mathrm{HO}_T(\cdot)$. Consequences for cross-validation will be derived in section 5.2.

3 Hypotheses

Approximating the process $HO_T(k)$, which is a sum over Fourier coefficients, inevitably involves bounding "tail sums" of the Fourier series, such as $\sum_{j=k+1} \theta_j^2$. To control these, we assume that the Fourier coefficients decay fast enough.

Hypothesis 1. *There exists constants $c_1 \geq 0$ and $\delta_1 \geq 0$ such that for all $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \leq \frac{c_1}{k^{2+\delta_1}}$.*

This upper bound is satisfied for some c_1, δ_1 if and only if the smoothness assumption $s \in H^\beta$ holds for some $\beta > 1$, where H^β denotes the Sobolev Hilbert space. Since we seek to approximate the discrete process $HO_T(k)$ by a continuous one, it is necessary to exclude the case where the bias decreases very fast, otherwise cross-validation is effectively performed over a finite set of very small models. For technical reasons, we will assume a polynomial growth rate, using the following three hypotheses.

Hypothesis 2. *There exists constants $c_2 \geq 0$, $\delta_2 \geq 0$ such that for all $k \in \mathbb{N}$, $\sum_{j=k+1}^{+\infty} \theta_j^2 \geq \frac{c_2}{k^{\delta_2}}$*

This hypothesis states that the Fourier coefficients θ_j^2 cannot decay faster than polynomially, and excludes in particular analytic functions. Hypothesis 2 holds for example if s or one of its derivatives has a point of discontinuity.

Hypothesis 3. *There exists constants $c_3 > 0$, $\delta_3 \geq \delta_6 > 0$ such that for all $k \geq 1$,*

$$\theta_k^2 \leq c_3 k^{-\delta_6} \sum_{j=0}^{\lfloor c_3 k^{\delta_3} \rfloor} \theta_{k+j}^2.$$

Hypothesis 3 means that the sequence θ_j^2 cannot have terms which are "much larger" than their neighbours to the right. Without such an assumption, the behaviour of the process may be dominated by a few large coefficients, leading to discontinuities in the mean or covariance functions. Together, hypotheses 1, 2, 3 basically mean that the Fourier coefficients of s on the cosine basis decrease polynomially. Let us now give a simpler condition which implies all of the previous assumptions.

Hypothesis 4. *There exists a slowly varying function L, a rate $\beta > 1$ and a constant C such that for all $j \in \mathbb{N}$,*

$$\theta_j^2 \leq C L(j) j^{-(2\beta+1)}$$

and such that one of the following equivalent conditions holds:

- *For some constant $\mu > 0$, the set*

$$J_\mu = \left\{ j \in \mathbb{N} : \theta_j^2 \geq \mu L(j) j^{-(2\beta+1)} \right\}$$

has positive lower density, i.e

$$\liminf_{k\to+\infty} \frac{|\{1, \dots, k\} \cap J_\mu|}{k} > 0.$$

- *There exists a constant* $0 < \mu_0$ *such that for all* $k \in \mathbb{N}$,

$$\sum_{j=k+1}^{+\infty} \theta_j^2 \geq \mu_0 L(k) k^{-2\beta}$$

Hypothesis 4 is a mild regularity condition on the coefficients θ_j, which requires coefficients to decrease no slower than a given rate, and to match this rate on a non-negligible subset of the integers. This excludes lacunary series, for example. Let us see how this assumption relates to the previous ones.

Lemme 3.1. *Under hypothesis 4, hypothesis 1 holds for any* $\delta_1 < 2(\beta-1)$, *hypothesis 2 holds for any* $\delta_2 < 2\beta$ *and hypothesis 3 holds with* $\delta_3 = \delta_6 = 1$.

Lemma 3.1 is proved in section 1 of the supplementary material. The two remaining hypotheses 5 and 6 do not bear on s, but on the parameter n_t which is chosen by the statistician.

Hypothesis 5. *There exists a constant* $\delta_4 > 0$ *such that* $n - n_t \leq n^{1-\delta_4}$.

This upper bound on the size of the validation sample $D_n^{T^c}$ ensures that there is enough noise in the estimate $\mathrm{HO}_T(\cdot)$ to guarantee a non-trivial limit process. For technical reasons, it is also necessary to lower bound $n - n_t$, which is accomplished using hypothesis 6 below.

Hypothesis 6. *There exists a constant* $\delta_5 > 0$ *such that* $n_v = n - n_t \geq n^{\frac{2}{3}+\delta_5}$.

The statistician can always choose n_t such that hypotheses 5 and 6 hold. One should however check that this is compatible with good performance of the hold-out. Oracle inequalities [2] show that the risk of the hold-out in model selection for L^2 density estimation is (at most) of order $\frac{n}{n_t}\mathrm{or}(n) + \frac{\log(n-n_t)}{n-n_t}$. If $\mathrm{or}(n)$ decreases in n with rate $\frac{1}{n^\alpha}$ ($\alpha \in \left(\frac{2}{3}, 1\right)$), which is the case under assumptions 1 and 2, $n - n_t$ can be chosen within the interval $[\frac{1}{2}n^{\frac{3\alpha+1}{4}}; n^{\frac{4+\alpha}{5}}]$ —so that assumptions 5 and 6 are satisfied— without changing the order of magnitude of the risk. Moreover, the optimal value of $n - n_t$ according to the above risk bound is of order $n^{\frac{1+\alpha}{2}}$ and hence belongs to the interval $[\frac{1}{2}n^{\frac{3\alpha+1}{4}}; n^{\frac{4+\alpha}{5}}]$ for n large enough.

4 Definitions

If a sequence k_n of model indices is not too badly chosen (such that $k_n \to +\infty$ and $\frac{k_n}{n} \to 0$), then the sequence of estimators $\hat{s}_{k_n}$ is consistent in L^2, which implies (under sufficient moment assumptions on s), that

$$\mathrm{HO}_T(k_n) \approx \left\|\hat{s}_{k_n}^T - s\right\|^2 - \|s\|^2 - 2(P_n^{T^c} - P)(s),$$

where the random error $(P_n^{T^c} - P)(s)$ is now independent of the model parameter k_n. While this error term is unavoidable when estimating the risk of a single estimator, when *comparing* several different models $k_{1,n}, \ldots, k_{j,n}$ in order to perform *model selection*, the quantities that matter are the pairwise differences, $\mathrm{HO}_T(k_{i,n}) - \mathrm{HO}_T(k_{j,n})$, in which the leading error term, $(P_n^{T^c} - P)(s)$, cancels out. To account for this effect, we consider a centered and scaled process of the following generic form:

$$\begin{aligned}\hat{R}_T^{ho}(\alpha) = \frac{1}{\mathfrak{e}}\left(\mathrm{HO}_T(k_n + \alpha\Delta) - \mathrm{HO}_T(k_n)\right) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}_{k_n+\alpha\Delta}^T - s\right\|^2 - \left\|\hat{s}_{k_n}^T - s\right\|^2\right) \\ - \frac{2}{\mathfrak{e}}(P_n^{T^c} - P)(\hat{s}_{k_n+\alpha\Delta}^T - \hat{s}_{k_n}^T),\end{aligned} \tag{3}$$

where $k_n, \Delta = \Delta(n)$ and $\mathfrak{e} = \mathfrak{e}(n)$ are sequences depending on $(n_t(n))_{n\in\mathbb{N}}$ and the density s.

4.1 Centering sequence

While any centering sequence $(k_n)_{n\in\mathbb{N}}$ may a priori be chosen, they are not equally interesting. With the goal of *model selection* in mind, we want the localized process (3) to provide information on the asymptotic location of $\operatorname{argmin}_{k\in\mathbb{N}}\{\mathrm{HO}_T(k)\}$ relative to the deterministic centering sequence k_n. This requires $\hat{R}_T^{ho}$ to be *coercive* as a function of α, that is to say

$$\lim_{\alpha\to\pm\infty} \hat{R}_T^{ho}(\alpha) = +\infty,$$

so that $\frac{\hat{k}_n - k_n}{\Delta}$ remains asymptotically bounded for any sequence $\hat{k}_n \in \operatorname{argmin}_{k\in\mathbb{N}}\{\mathrm{HO}_T(k)\}$. While this may not always be achievable (the risk may have distant global minima), it is at least possible to guarantee that $\mathbb{E}\left[\hat{R}_T^{ho}(\alpha)\right] \geq 0$ for all α, by choosing

$$k_n \in \operatorname*{argmin}_{k\in\mathbb{N}} \left\{\mathbb{E}\left[\mathrm{HO}_T(k)\right]\right\}. \tag{4}$$

In order to obtain a more explicit formula for the centering sequence k_n, we slightly modify this definition by replacing $\mathbb{E}\left[\mathrm{HO}_T(k)\right]$ by an approximation. Consider the following definition.

Definition 4. *For all $n \in \mathbb{N}$, let*

$$\mathrm{or}(n) = \min_{k\in\mathbb{N}} \left\{ \sum_{j=k+1}^{+\infty} \theta_j^2 + \frac{k}{n} \right\}$$

$$\textit{and} \qquad k_*(n) = \max \operatorname*{argmin}_{k\in\mathbb{N}} \left\{ \frac{k}{n} + \sum_{j=k+1}^{+\infty} \theta_j^2 \right\}.$$

This definition is based on the following approximation to the L^2 risk, a precise statement of which can be found in claim 5:

$$\left\|\hat{s}_k^T - s\right\|^2 \approx \mathbb{E}\left[\left\|\hat{s}_k^T - s\right\|^2\right] = \mathbb{E}\left[\mathrm{HO}_T(k)\right] \approx \frac{k}{n_t} + \sum_{j=k+1}^{+\infty} \theta_j^2.$$

This leads to the choice $k_n = k_*(n_t(n))$, which we will often denote simply by k_*.

4.2 Scaling

It remains now to choose the scaling sequences $\Delta(n), \mathfrak{e}(n)$. Since $\mathfrak{e}(n)$ will be chosen so as to obtain a bounded, non-vanishing process, its value is essentially determined by that of $\Delta(n)$. Let us now discuss how this sequence is defined. It has already been remarked that, provided that the sequence $n_t(n)$ is properly chosen, the hold-out risk estimator, $\mathrm{HO}_T(k)$, concentrates around the risk of $\hat{s}_k^T$, which itself concentrates around its expectation. Thus, if the scale Δ is too large, the centered and rescaled hold-out process (equation (3)) will converge to a deterministic function. The scale chosen is then too large to distinguish the hold-out estimator from the quantity that it estimates. As the scale decreases, the effect of the random fluctuations is magnified and at a certain critical scale, randomness should appear in the asymptotic. Our goal is now to characterize this critical scale, as well as the corresponding limit process.

An appropriate choice of $\Delta, \mathfrak{e}$ is given in the following definition.

Definition 5. *For all $n \in \mathbb{N}$, let*

$$\Delta_d(s,n_t,n) = \max\left\{ l \in \mathbb{N} : \frac{1}{l}\sum_{j=1}^{l} \theta^2_{k_*(n_t)+j} \geq \left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t} \right\}$$

$$\Delta_g(s,n_t,n) = \max\left\{ l \in \{0,\dots,k_*(n_t)\} : \frac{1}{l}\sum_{j=0}^{l-1} \theta^2_{k_*(n_t)-j} \leq \left[1 + \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{l}}\right]\frac{1}{n_t} \right\}$$

$$\Delta(s,n_t,n) = \max\left(\Delta_d(s,n_t,n), \Delta_g(s,n_t,n)\right)$$

$$\mathcal{E}(s,n_t,n) = \frac{\Delta(s,n_t,n)}{n_t}.$$

$$\mathfrak{e}(s,n_t,n) = \sqrt{\frac{\mathcal{E}(s,n_t,n)}{n-n_t}}.$$

As the sequence $n_t(n)$ and the density s are considered to be fixed once and for all, the notation

$$\Delta(s,n_t,n), \mathcal{E}(s,n_t,n), \mathfrak{e}(s,n_t,n)$$

will frequently be replaced by the abbreviations $\Delta, \mathcal{E}, \mathfrak{e}$.

4.3 Rescaled process

Now that $\Delta, \mathfrak{e}$ are defined, the hold-out process can be rescaled as in equation (6). More precisely, the rescaled hold-out process is given by Definition 6 below.

Definition 6. *For all $j \in [-k_*; +\infty[\cap\mathbb{Z}$, let*

$$\hat{R}^{ho}_T\left(\frac{j}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left(HO_T\left(k_*+j\right) - HO_T\left(k_*\right)\right),$$

in other words (by definition 2)

$$\hat{R}^{ho}_T\left(\frac{j}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2\right) - \frac{2}{\mathfrak{e}}\left(P^{T^c}_n - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right).$$

The $\hat{R}^{ho}_T$ function is extended by linear interpolation to all $\alpha \in \left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$. Let $\hat{R}^{cv}_{\mathcal{T}}$ be defined in a similar manner, i.e

$$\hat{R}^{cv}_{\mathcal{T}}(\tfrac{j}{\Delta}) = \frac{1}{\mathfrak{e}}\left(\mathrm{CV}_{\mathcal{T}}(k_*+j) - \mathrm{CV}_{\mathcal{T}}(k_*)\right)$$

for all $j \in [-k_*; +\infty[\cap\mathbb{Z}$, extended by linear interpolation to $\left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$.

Note that by linearity of the interpolation operation, $\hat{R}^{cv}_{\mathcal{T}} = \frac{1}{V}\sum_{i=1}^{V}\hat{R}^{ho}_{T_i}$. The extension of $\hat{R}^{ho}_T, \hat{R}^{cv}_{\mathcal{T}}$ by linear interpolation simplifies their approximation by a continuous process. Notice that any minimizer of $\hat{R}^{ho}_T$ (resp. $\hat{R}^{cv}_{\mathcal{T}}$) on the grid $\frac{1}{\Delta}([-k_*(n_t); +\infty[\cap\mathbb{Z})$ remains a minimizer of $\hat{R}^{ho}_T$ (resp. $\hat{R}^{cv}_{\mathcal{T}}$) on the interval $\left[\frac{-k_*(n_t)}{\Delta}; +\infty\right[$. In particular, this applies to the hold-out parameter obtained by minimisation of the hold-out risk estimator.

4.4 Mean function and domain of approximation

The following function is approximately the mean function of the random process $\hat{R}_T^{ho}(\cdot)$. It plays an essential role in the results of this article.

Definition 7. *For all $k \in \mathbb{N}$, let $R(k) = \sum_{j=k+1}^{+\infty} \theta_j^2$. Extend R to $\mathbb{R}_+$ by linear interpolation:*

$$\forall x \in \mathbb{R}_+, R(x) = (1 + \lfloor x \rfloor - x)R(\lfloor x \rfloor) + (x - \lfloor x \rfloor)R(\lfloor x \rfloor + 1).$$

$f_n :]-\frac{k_*(n_t)}{\Delta}; +\infty[\to \mathbb{R}_+$ is now defined by:

$$f_n(\alpha) = \frac{1}{\mathfrak{e}} \left(R(k_*(n_t) + \alpha\Delta) - R(k_*(n_t)) + \frac{\alpha\Delta}{n_t} \right). \tag{5}$$

Thus, for all $k \in \mathbb{N}$, $k \neq k_*(n_t)$,

$$\mathfrak{e} f_n \left(\frac{k - k_*(n_t)}{\Delta} \right) = \left| \sum_{j=k \wedge k_*(n_t)+1}^{k \vee k_*(n_t)} \theta_j^2 - \frac{1}{n_t} \right|. \tag{6}$$

It is clear by equation (6) that f_n reaches its (global) minimum at 0, moreover f_n is the sum of a non-increasing function and a non-decreasing linear function. For technical reasons, our methods do not let us carry out approximation over the whole domain. In order to limit the domain of approximation in a principled manner, we introduce the following definition.

Definition 8. *Let $T \subset \{1 \dots n\}$ be a subset of cardinality n_t and let $k_* = k_*(n_t)$. Given any $x > 0$, let*

$$a_x = \min \left\{ \alpha \in \frac{1}{\Delta}\mathbb{Z}, \mathfrak{e} f_n(\alpha) \leq x \left[2or(n_t) + \frac{1}{n - n_t} \right] \right\}$$
$$b_x = \max \left\{ \alpha \in \frac{1}{\Delta}\mathbb{Z}, \mathfrak{e} f_n(\alpha) \leq x \left[2or(n_t) + \frac{1}{n - n_t} \right] \right\}.$$

We shall see later that not much information is lost by restricting the process to the interval $[a_x; b_x]$.

5 Main theorems

Let us now give simple approximants $Y_{n,V}(u)$ to the rescaled CV estimators $\hat{R}_T^{cv}(u)$ (including the hold-out when the number of splits $V = 1$).

5.1 Simple validation

The following theorem shows that the process $\hat{R}_T^{ho}(\cdot)$ can be approximated on $[a_x, b_x]$ by the sum of f_n and a time-changed Brownian motion.

Theorem 1. *Assume that the hypotheses of section 3 hold. There exists an increasing continuous function $g_n : [-\frac{k_*(n_t)}{\Delta}; +\infty[\to \mathbb{R}$ and for any $x > 0$, there exists a two-sided Brownian motion $(W_t)_{t \in [a_x; b_x]}$ independent from D_n^T such that, for any $y > 0$, with probability greater than $1 - e^{-y}$,*

$$\mathbb{E}\left[\sup_{u \in [a_x; b_x]} \frac{\left| \hat{R}_T^{ho}(u) - (f_n(u) - W_{g_n(u)}) \right|}{1 + f_n(u)} \,\middle|\, D_n^T \right] \leq \kappa_0 (1+y)^2 (1+x)^{u_6} \left(n^{-u_1} + (\log n)^{\frac{1}{6}} \left(\frac{k_*^{\delta_3 - \delta_6}}{\Delta} \right)^{\frac{1}{36}} \left(\frac{k_*^{\delta_3 - \delta_6}}{\Delta} \vee 1 \right)^{\frac{1}{18}} \right), \tag{7}$$

where $u_1 > 0, u_6 \geq 0$ and $\kappa_0 \geq 0$ are constants which depend only on $(\delta_i)_{1\leq i\leq 6}$ and c_1, c_2, δ_1, c_3, respectively. Moreover, g_n and W can be chosen so as to satisfy the following conditions.

1. $g_n(0) = 0$, $W_0 = 0$,
2. $\forall(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2$, $\alpha_2 < \alpha_1 \implies g_n(\alpha_1) - g_n(\alpha_2) \geq 4\|s\|^2 [\alpha_1 - \alpha_2]$.
3. *For all* $(\alpha_1, \alpha_2) \in \left[\frac{-k_*}{\Delta}; +\infty\right[^2$ *such that* $\alpha_1 < \alpha_2 < 0$ *or* $0 < \alpha_1 < \alpha_2$,

$$g_n(\alpha_2) - g_n(\alpha_1) \leq -\frac{8\|s\|_\infty}{(n-n_t)\mathfrak{e}}[f_n(\alpha_2) - f_n(\alpha_1)] + \left(8\|s\|_\infty + 4\|s\|^2\right)[\alpha_2 - \alpha_1]. \tag{8}$$

It states that the rescaled hold-out process, $\hat{R}_T^{ho}$, can be approximated on $[a_x, b_x]$ by a continuous process $Y_{n,1}$, which is the sum of the non-negative, deterministic function f_n and a time-changed Brownian motion W_{g_n}. If $k_*^{\delta_3-\delta_6} = o(\Delta)$, the error made in this approximation is negligible relative to $1 + f_n$. A sufficient (but not necessary) condition for this is that $\delta_3 - \delta_6 < \delta_4$, where δ_4 is the constant introduced in hypothesis 5. In particular, hypothesis 4 is sufficient since it yields $\delta_3 = \delta_6 = 1$.

f_n and g_n depend on n_t and n, but not on the data (they are deterministic functions), while W depends on the data only through the test sample $D_n^{T^c}$. In particular, in this asymptotic setting, $\hat{R}_T^{ho}$ doesn't depend on D_n^T, the training data. Lemma 5.1 below gives upper and lower bounds on f_n which prove the non-triviality of Theorem 1. Its proof can be found in appendix A.2.

Lemme 5.1. *There exists some $t \in \{-1, 1\}$ such that $f_n(t) \leq 1$. In other words,*

$$\min_{t\in\{-1,1\}} f_n(t) \leq 1.$$

Moreover, for all $l \in \mathbb{Z}$ such that $|l| > \Delta$,

$$f_n\left(\frac{l}{\Delta}\right) \geq \sqrt{\frac{|l|}{\Delta}}$$

and for all $\alpha \in \mathbb{R}$,

$$f_n(\alpha) \geq \sqrt{|\alpha|} - 1.$$

Thus, f_n is upper bounded by 1 at some point $t_n \in \{-1; 1\}$. The lower bound, lemma 5.3 below and Definition 8 imply that for $x \geq 1$, $[a_x, b_x]$ either contains $I_n = [0, 1]$ (if $t_n = 1$) or $I_n = [-1, 0]$ (if $t_n = -1$), the two cases being non-exclusive. Moreover, by equation (8), g_n is bounded on this interval $I_n \in \{[0; 1], [-1; 0]\}$ and its total variation on I_n depends only on $\|s\|^2, \|s\|_\infty$. On the other hand, by point 2 of Theorem 1, the function g_n increases on its domain at least as fast as the linear function $\alpha \mapsto 4\|s\|^2 \alpha$. In particular, f_n, g_n are both bounded at t_n and moreover, $g_n(t_n) \geq 4\|s\|^2$. It follows that $f_n(t_n) - W_{g_n(t_n)}$ is a tight sequence of Gaussian random variables with non-vanishing variance, which proves the presence of randomness in the asymptotic. If f_n is convex on $[a_x; b_x]$ (which corresponds to the sequence of squared coefficients θ_j^2 being non-increasing on ΔI_n), f_n is also 2−Lipschitz on $\frac{1}{2}I_n$ (say) and it follows that g_n is Lipschitz-continuous on $\frac{1}{2}I_n$. Figure 1 illustrates the bounds that hold on f_n, g_n in this case, assuming $\Delta = \Delta_d$.

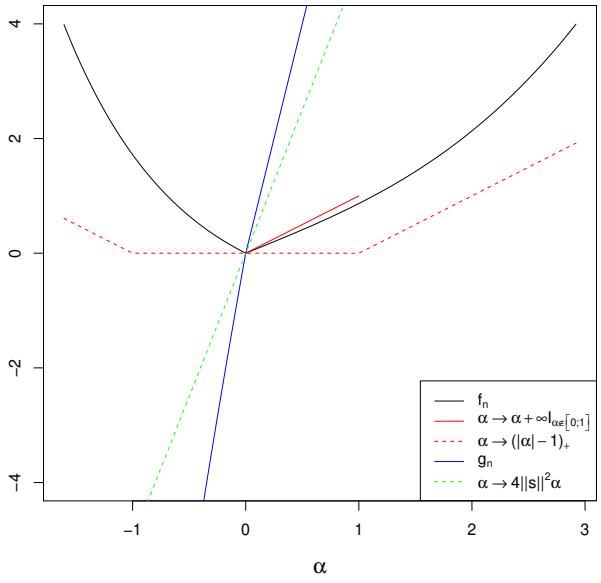

Figure 1: A plot of f_n, g_n on $[a_6; b_6]$ with upper and lower bounds, with a convex f_n and $\|s\|^2 = 1.2$.

5.2 Incomplete $V-$fold cross-validation

Since the cross-validation risk estimator $\mathrm{CV}_{\mathcal{T}}(k)$ can be written as an average of hold-out risk estimators $\mathrm{HO}_{T_i}(k)$, Theorem 1 has direct implications for CV.

Corollary 2. *Assume that the hypotheses of section 3 hold. Let f_n, g_n be as in definition 7 and theorem 1. For any $x > 0$,*

$$\mathbb{E}\left[\sup_{u\in[a_x;b_x]} \frac{\left|\hat{R}^{cv}_{\mathcal{T}}(u) - (f_n(u) - W_{\frac{g_n(u)}{V}})\right|}{1+f_n(u)}\right] \leq \kappa_0(1+x)^{u_6}\left(n^{-u_1} + (\log n)^{\frac{1}{6}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\right)^{\frac{1}{36}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\vee 1\right)^{\frac{1}{18}}\right),$$

with the same constants κ_0, u_1 as in Theorem 1.

Proof. By integrating the bound of Theorem 1,

$$\mathbb{E}\left[\sup_{u\in[a_x;b_x]} \frac{\left|\hat{R}^{ho}_{T_i}(u) - (f_n(u) - W^i_{\frac{g_n(u)}{V}})\right|}{1+f_n(u)}\right] \leq \kappa_0(1+x)^{u_6}\left(n^{-u_1} + (\log n)^{\frac{1}{6}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\right)^{\frac{1}{36}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\vee 1\right)^{\frac{1}{18}}\right),$$

for each $i \in \{1,\dots,V\}$, where the W^i are symmetrical BMs that are independent of $D_n^{T_i}$. We can construct W^i such that $W^i = H(D_n^{T_i^c}, U_i)$, where H is a measurable function and U_i is an auxiliary uniform random variable. Taking independent U_i yields i.i.d W^i, since the sets $T_i^c = I_i$ are disjoint. Let $\bar{W} = \frac{1}{V}\sum_{i=1}^V W^i$. By Jensen's inequality,

$$\mathbb{E}\left[\sup_{u\in[a_x;b_x]} \frac{\left|\hat{R}^{cv}_{\mathcal{T}}(u) - (f_n(u) - \bar{W}_{g_n(u)})\right|}{1+f_n(u)}\right] \leq \kappa_0(1+x)^{u_6}\left(n^{-u_1} + (\log n)^{\frac{1}{6}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\right)^{\frac{1}{36}}\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\vee 1\right)^{\frac{1}{18}}\right).$$

Conclude by noting that $(\bar{W}_t)_{t\in\mathbb{R}}$ is equal in distribution to $(W_{t/V})_{t\in\mathbb{R}}$, as a continuous random process. □

Corollary 2 proves that cross-validation is effective at reducing the variance of risk estimation, compared to simple validation (the hold-out). The process approximating $\hat{R}^{cv}_{\mathcal{T}}$ is of the same form as that approximating $\hat{R}^{ho}_T$, but with its variance reduced by a factor V, as would be the case if the hold-out estimators $\mathrm{HO}_{T_i}(\cdot)$ were independent. Importantly, this reduction in variance occurs for the rescaled process $\hat{R}^{cv}_{\mathcal{T}}$, and so can be expected to reflect the model selection performance. Corollary 2 is sharp when V is fixed as $n \to +\infty$, since in that case, the approximating process $f_n - W_{g_n/V}$ remains nontrivial (random and of bounded size) as $n \to +\infty$. When $V = V_n \to +\infty$, corollary 2 implies that $\hat{R}^{cv}_{\mathcal{T}}$ concentrates around the deterministic function f_n. If the sequence V_n grows slowly enough, such that $V_n = o(n^{u_1})$, corollary 2 also provides the rate of this convergence (which is $\frac{1}{\sqrt{V_n}}$).

5.3 Larger scales

Theorem 1 leaves open the question of what happens outside the window $[a_x, b_x]$. For a point $t = \frac{j}{\Delta}$ to lie outside this interval, the excess risk of $\hat{s}_k^T$, which is approximately $\mathfrak{e}f_n(t)$, must exceed the optimum by a constant factor x. Known consistency results for cross-validation [2] show that the minimum of $\hat{R}^{cv}_{\mathcal{T}}$ lies within such intervals with high probability. This suggests that the random fluctuations of the process are irrelevant outside $[a_x; b_x]$. The following proposition shows more precisely that $\hat{R}^{cv}_{\mathcal{T}}$ concentrates around the deterministic function f_n as $x \to +\infty$.

Proposition 5.2. *Let $(M_n)_{n\in\mathbb{N}}$ be an integer sequence and*

$$I_n = \left[-\frac{k_*(n_t)}{\Delta}, \frac{M_n - k_*(n_t)}{\Delta}\right] \cap \frac{1}{\Delta}\mathbb{Z}.$$

Assume that there exist constants $\delta_7 > 0, c_7$ such that

$$\frac{k_*^{\delta_3-\delta_6}}{\Delta} \leq c_7 n^{-\delta_7}$$

for all integers n. There exists constants κ and $\delta > 0$ (depending on $(c_i, \delta_i)_{i\in\{1,\dots,7\}}$) such that, for all $x \geq e^2$,

$$\mathbb{E}\left[\sup_{t\in I_n\setminus[a_x,b_x]}\left|\frac{\hat{R}_{\mathcal{T}}^{cv}(t)}{f_n(t)} - 1\right|\right] \leq \sqrt{\frac{2\kappa\log\log(\kappa x)}{Vx}} + \frac{\kappa\log^2(nM_n)}{n^\delta}.$$

The proof of proposition 5.2 can be found in section 3 of the supplementary material.Because a union bound is used, proposition 5.2 applies only on an interval I_n corresponding to values of the dimension parameter $k \leq M_n$. M_n can grow polynomially or even super-polynomially without invalidating the result. In practice, this assumption is mild since by definition 4, $k_*(n_t) \leq \|s\|^2 n_t \leq n^2$ for any $n \geq \|s\|^2$. Figure 2 gives an illustration of the situation for $x = 25$ and

$$f_n : \alpha \mapsto \begin{cases} e^{-\alpha} - 1 \text{ if } \alpha \leq 0 \\ \frac{8}{10}\alpha + \frac{8}{30}\alpha^3 \text{ if } \alpha \geq 0 \end{cases}$$

$$g_n : \alpha \mapsto \begin{cases} 7.8\alpha \text{ if } \alpha \geq 0 \\ 7.8\alpha - 3f_n(\alpha) \text{ if } \alpha \leq 0 \end{cases}$$

(which satisfy the properties of lemma 5.1 and Theorem 1 when $\|s\|^2 \leq 1.2$ and $\|s\|_\infty \leq 1.5$). Figure 2 shows how for large x, $\sqrt{g_n}$ and hence W_{g_n} become negligible compared to f_n outside the interval $[a_x, b_x]$, as stated by proposition 5.2. Hence, for sufficiently large x, the random term W_{g_n} becomes negligible relative to the deterministic f_n outside the interval $[a_x; b_x]$.

5.4 Bounds and estimates for $\Delta, \mathfrak{e}$

$\Delta = \Delta(s, n_t, n)$ and $\mathfrak{e}$ depend in a complicated way on the density s through its Fourier coefficients $(\theta_j)_{j\in\mathbb{N}}$. Thus, it is not intuitively clear how large $\Delta, \mathcal{E}$ and $\mathfrak{e}$ are. The following lemma gives some simple universal bounds on these quantities.

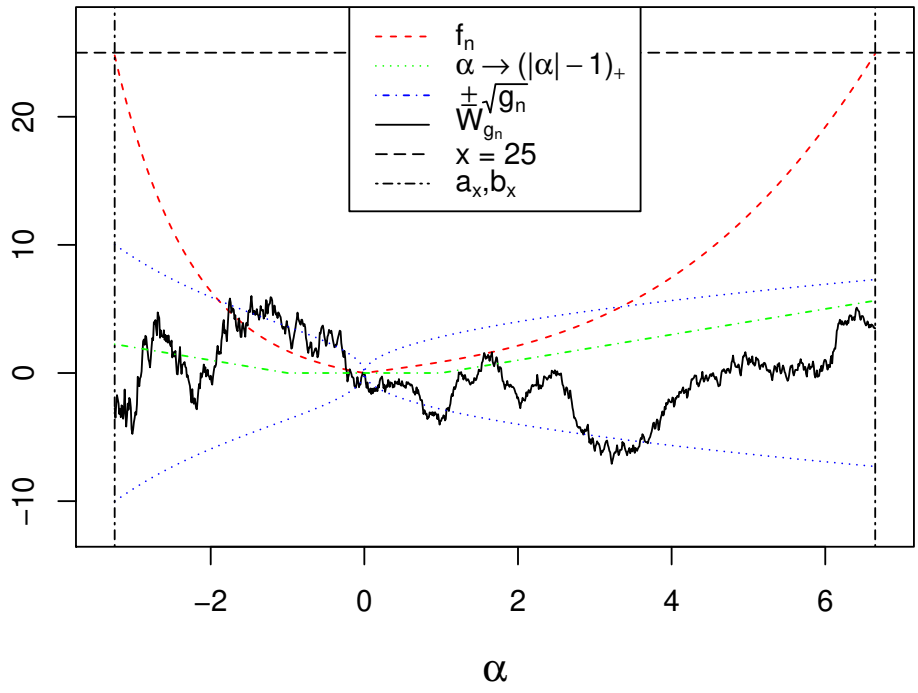

Figure 2: A plot of f_n, W_{g_n} on $[a_x; b_x]$, for $x = 25$, $g_n : \alpha \mapsto 7.8\alpha - 3f_n(\alpha)\mathbb{I}_{\alpha<0}$.

Lemme 5.3. *For any density s,*

$$\Delta \geq \frac{n_t}{n - n_t} \tag{9}$$

$$\mathcal{E} \geq \frac{1}{n - n_t} \tag{10}$$

$$\mathfrak{e} \geq \frac{1}{n - n_t} \tag{11}$$

$$\mathfrak{e} \leq \mathcal{E} \tag{12}$$

$$\mathcal{E} \leq 2\mathrm{or}(n_t) + \frac{1}{n - n_t}. \tag{13}$$

This lemma is proved in appendix A.1. The first three equations are equivalent and imply that the size of the random fluctuations of $\mathrm{HO}_T(k)$ are at least of order $\frac{1}{n-n_t}$ when localized around $k_*(n_t)$, i.e the inverse of the size of the test sample. On the other hand, equation (13) implies an upper bound of

$$\mathfrak{e} \leq \sqrt{\frac{2\mathrm{or}(n_t)}{n - n_t}} + \frac{1}{n - n_t},$$

which involves only the size of the test set $n - n_t$ and the minimal (oracle) risk over the model collection, $\mathrm{or}(n_t)$. This upper bound is negligible relative to $\mathrm{or}(n_t)$ if $\mathrm{or}(n_t) \gg \frac{1}{n-n_t}$. Hence, $\mathfrak{e}$ is negligible relative to $\mathrm{or}(n_t)$ if and only if $\mathrm{or}(n_t) \gg \frac{1}{n-n_t}$. In that case, the random fluctuations of $\mathrm{HO}_T(\cdot)$ are asymptotically negligible around $k_*(n_t)$, which suggests risk consistency of hold-out model selection. This can be compared with known oracle inequalities for simple validation which imply risk consistency under the slightly stronger assumption that $\mathrm{or}(n_t) \gg \frac{\log(n_t)}{n-n_t}$. The following two examples show that these bounds are tight, at least up to constants.

Example 5.1. *Let u_n be an integer sequence such that $u_n \to +\infty$ and $u_n \leq \frac{\sqrt{n}}{2}$ for all n. Assume that $\frac{n_t(n)}{n} \to 1$ and*

$$\frac{1}{n - n_t(n)} = o\left(\frac{u_{n_t(n)}}{n_t(n)}\right).$$

Define the sequence of pdfs

$$s_n = 1 + \sum_{j=1}^{u_{n_t(n)}} \sqrt{\frac{1}{n_t(n)}}\psi_j.$$

Then $\mathcal{E}(s_n, n_t(n), n) \sim \frac{u_{n_t(n)}}{n_t(n)} \sim \mathrm{or}(s_n, n_t(n))$.

Proof. Remark that definition 4 implies that $k_*(n_t) = u_{n_t}$. Then as $n \to +\infty$, $\mathcal{E}(s_n, n_t, n) \sim \frac{u_{n_t}}{n_t} \sim \mathrm{or}(n_t)$ and $\frac{n-n_t}{n_t} = o(\mathrm{or}(n_t))$, so $\mathfrak{e}(s_n, n_t, n) = o(\mathcal{E}(s_n, n_t, n))$. □

Example 5.2. *Define the pdf*

$$s = \sum_{j=0}^{+\infty} \frac{\psi_j}{3^j}.$$

Assume that the sequence $n_t(n)$ is such that $\frac{n_t(n)}{n} \to 1$. Then $\Delta(s, n_t(n), n) \sim \frac{n_t(n)}{n-n_t(n)}$ as $n \to +\infty$.

Proof. The Fourier coefficients of s are

$$\forall j \in \mathbb{N}, \langle s, \psi_j \rangle = \theta_j = \frac{1}{3^j}. \tag{14}$$

Since the sequence θ_j^2 is non-increasing,

$$k_*(n_t) = \max\left\{k \geq 1 : \frac{1}{9^k} \geq \frac{1}{n_t}\right\}.$$

Then by Lemma 5.3, $\Delta_d \geq \frac{n_t}{n-n_t}$, but

$$\frac{1}{\Delta_d}\sum_{j=1}^{\Delta_d} \theta_{k_*+j}^2 \leq \frac{1}{\Delta_d}\sum_{j=1}^{\Delta_d} \frac{1}{9^j n_t} \sim \frac{9}{8n_t}\frac{1}{\Delta_d} = o\left(\frac{1}{n_t}\right).$$

It follows by definition of Δ_d that

$$\left(1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}}\right) \leq o(1),$$

which yields $\Delta_d(s, n_t, n) \sim \frac{n_t}{n-n_t}$. Moreover, by definition for any $l \leq k_*$

$$\begin{aligned}\frac{1}{l}\sum_{j=0}^{l-1} \theta_{k_*-j}^2 &\geq \frac{1}{l}\sum_{j=0}^{l-1} \frac{9^j}{n_t} \\ &\geq \frac{9^l}{8ln_t}.\end{aligned}$$

Hence, if Δ_g were greater than $\frac{n_t}{n-n_t} + 1$, we would have that

$$\frac{9^{\Delta_g}}{8\Delta_g} \leq 2,$$

which is impossible for large enough $\frac{n_t}{n-n_t}$. Thus,

$$\Delta = \max\left(\Delta_g, \Delta_d\right) \sim \frac{n_t}{n-n_t}.$$

□

The bounds of lemma 5.3 do not fully determine the order of magnitude of $\Delta, \mathfrak{e}$: in general, the ratio between the upper and lower bounds is a power of n. In contrast, direct statistical applications of Theorem 1, such as confidence bands for the risk, require Δ and $\mathfrak{e}$ to be known up to a constant factor. To this end, we now investigate the possibility of estimating $\Delta, \mathfrak{e}$ from the data. In this section, estimators of $\hat{\Delta}$ are constructed which are provably consistent (up to a constant factor) under the assumption that the coefficients θ_j^2 are non-increasing (which corresponds to a convexity assumption on R and f_n). We believe that this assumption may be further relaxed, at least to approximate convexity of f_n (i.e $f_n = \overline{f}_n + o(\mathfrak{e})$ for some convex $\overline{f}_n$), though for simplicity's sake we do not pursue this refinement here. It seems however that some assumption of this kind is necessary to prevent the maximum in the definition of Δ_d, Δ_g from being unduly sensitive to small perturbations of the coefficients.

We first construct an estimator of $\hat{\Delta}$ based on a preliminary estimate of $k_*(n_t)$ (for example, the hold-out risk minimizer). We will later show how to dispense with this preliminary estimator, at the price of a possible loss of accuracy.

Let K_n be an integer sequence growing at a polynomial rate and such that $K_n \geq k_*(n_t) + \Delta$ for all n large enough (for example, $K_n = n$). For any $k \in \{1, \ldots, K_n\}$, let

$$\hat{\Delta}_d(k) = \max \left\{ l \in \{1, \ldots, K_n - k\} : \frac{1}{l} \sum_{j=1}^{l} \left(\hat{\theta}^T_{\hat{k}_*(n_t)+j}\right)^2 \geq \left[2 - \sqrt{\frac{n_t}{n - n_t}} \frac{1}{\sqrt{l}}\right] \frac{1}{n_t} \right\} \tag{15}$$

$$\hat{\Delta}_g(k) = \max \left\{ l \in \{1, \ldots, k\} : \frac{1}{l} \sum_{j=0}^{l-1} \left(\hat{\theta}^T_{\hat{k}_*(n_t)-j}\right)^2 \leq \left[2 + \sqrt{\frac{n_t}{n - n_t}} \frac{1}{\sqrt{l}}\right] \frac{1}{n_t} \right\} \tag{16}$$

Finally, let $\hat{\Delta}(k) = \max\left(\hat{\Delta}_g(k), \hat{\Delta}_d(k)\right)$. Note the change in the constant on the right hand-side from 1 to 2: this is meant to compensate for the bias of $(\hat{\theta}^T_k)^2$ as an estimator of θ_k^2. The estimator $\hat{\Delta}(k)$ has the following property.

Theorem 3. *Assume that the sequence of Fourier coefficients $(\theta_j)_{j\geq 1}$ is non-increasing in absolute value. For any $k \in \{1, \ldots, K_n\}$, let*

$$\alpha_k = \frac{k - k_*(n_t)}{\Delta}.$$

Fix some $x > 0$. There exists constants $\kappa, u_8 > 0$ such that the following property obtains with probability greater than $1 - e^{-x}$ and for the sequence $\varepsilon_n = \kappa(1 + x + \log n)n^{-u_8}$.

For all integers n such that $K_n \geq k_* + \Delta$ and $\varepsilon_n \leq \frac{1}{10}$, for all $k \in \{1, \ldots, K_n\}$,

$$\hat{\Delta}(k) \leq \Delta \max\left(1 + |\alpha_k| + \varepsilon_n, \frac{1}{\left(1 - \frac{5}{2}\varepsilon_n\right)^2} + \frac{\varepsilon_n(2 + 5f_n(\alpha_k))}{1 - \frac{5}{2}\varepsilon_n}\right) \tag{17}$$

and

$$\hat{\Delta}(k) \geq \begin{cases} \frac{\Delta}{(1+2\varepsilon_n)^2} \max\left(\frac{1}{4}, \sqrt{1 - 4((1+\varepsilon_n)|\alpha_k| + \varepsilon_n(2+\varepsilon_n))}\right) & \text{if } |\alpha_k| \leq \frac{1}{4} - \frac{9}{4}\varepsilon_n \\ |\alpha_k|\Delta \text{ if } |\alpha_k| \geq \frac{1}{4} - \frac{9}{4}\varepsilon_n \geq 2\varepsilon_n^2 \end{cases} \tag{18}$$

Theorem 3 is proved in section 4 of the supplementary material. It shows that $\hat{\Delta}(k)$ is consistent for some (random) k provided that $|\alpha_k| \to 0$ and $f_n(\alpha_k) \to 0$, meaning that the error of k as an estimator of $k_*(n_t)$ is negligible at the scale Δ, in terms of absolute value and in terms of risk (the function f_n). Moreover, the ratio $\frac{\hat{\Delta}(k)}{\Delta}$ remains bounded and bounded away from 0 provided only that $|\alpha_k|$ and $f_n(\alpha_k)$ remains bounded. This is sufficient for practical purposes, since the precise value of Δ is not what matters: rather, it is the fact that renormalizing by Δ leads to a non-trivial asymptotic. This property is robust to perturbation of Δ by constant factors. We believe that the assumption "$|\alpha_k|$ and $f_n(\alpha_k)$ remain bounded" holds when k is the minimizer of the hold-out or cross-validated risk. However, we may dispense with this conjecture by choosing a different estimator:

Corollary 4. *Let K_n be a sequence of integers. Define*

$$\hat{\Delta} = \min\left\{\hat{\Delta}(k) : 1 \leq k \leq K_n\right\}.$$

With probability greater than $1-e^{-x}$*, for all integers* n *such that* $K_n \geq k_* + \Delta$ *and* $\varepsilon_n = \kappa(1+x)n^{-u_8} \leq \frac{1}{10}$*,*

$$\left(\frac{1-9\varepsilon_n}{4}\right)\Delta \leq \hat{\Delta} \leq \frac{1+2\varepsilon_n - 5\varepsilon_n^2}{\left(1-\frac{5}{2}\varepsilon_n\right)^2}\Delta.$$

Proof. Let

$$\hat{k}_* = \min \operatorname{argmin}\left\{\hat{\Delta}(k) : 1 \leq k \leq K_n\right\}.$$

and let $\hat{\alpha}_* = \frac{\hat{k}_* - k_*}{\Delta}$. By Theorem 3, we have that

$$\begin{aligned}\hat{\Delta} &= \min\left\{\hat{\Delta}(k) : 1 \leq k \leq K_n\right\} \\ &\leq \hat{\Delta}(k_*) \\ &\leq \Delta \max\left(1+\varepsilon_n, \frac{1}{(1-\frac{5}{2}\varepsilon_n)^2} + \frac{2\varepsilon_n}{1-\frac{5}{2}\varepsilon_n}\right).\end{aligned}$$

On the other hand, if $|\hat{\alpha}_*| \leq \frac{1-9\varepsilon_n}{4}$, we have that

$$\hat{\Delta} = \hat{\Delta}(\hat{k}_*) \geq \frac{1}{4(1+2\varepsilon_n)^2} \geq \frac{1}{4}(1-4\varepsilon_n) \geq \frac{1-9\varepsilon_n}{4}$$

while if $|\hat{\alpha}_*| \geq \frac{1-9\varepsilon_n}{4}$, then

$$\hat{\Delta} \geq |\hat{\alpha}_*|\Delta \geq \frac{1-9\varepsilon_n}{4}\Delta.$$

This proves the lower bound. □

Thus, Δ may be estimated up to an asymptotic constant factor of 2 under the assumption that f_n is convex. It follows that $\mathfrak{e}$ may likewise be estimated up to a constant factor of $\sqrt{2}$, by "plugging in" the estimator of Δ into the formula $\mathfrak{e} = \sqrt{\frac{\Delta}{n_t(n-n_t)}}$.

6 Proof sketch

In this section, we sketch the proof of the main theorem (Theorem 1), laying out the main intermediate results. The hold-out risk estimator can be expressed as the sum of two terms. Definition 9 below gives a name to each of these terms.

Definition 9. *For all* $j \in [-k_*(n_t); +\infty[\cap\mathbb{Z}$*, let*

$$L(\tfrac{j}{\Delta}) = \frac{1}{\mathfrak{e}}\left(\left\|\hat{s}^T_{k_*+j} - s\right\|^2 - \left\|\hat{s}^T_{k_*} - s\right\|^2\right).$$

The function L *is extended to the interval* $[-\frac{k_*(n_t)}{\Delta}; +\infty[$ *by linear interpolation. Let* Z *be the random function defined for all* $j \in [-\frac{k_*(n_t)}{\Delta}; +\infty[\cap\mathbb{Z}$ *by*

$$Z\left(\frac{j}{\Delta}\right) = \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right)$$

and extended by linear interpolation to the interval $[-\frac{k_*(n_t)}{\Delta}; +\infty[$*, so that for all* α*,* $\hat{R}^{ho}_T(\alpha) = L(\alpha) - Z_\alpha$*.*

Thus, L is the rescaled excess risk, and Z is a centered empirical process conditionally on D_n^T. These two terms will be approximated separately. First, L can be approximated by f_n using a concentration inequality [2, Lemma 14]. This leads to the following claim.

Claim 5. *Let L be the function introduced in definition 9, and f_n be given by definition 7. There exists a constant κ_1 such that, for any $x, y > 0$, with probability greater than $1 - e^{-y}$,*

$$\sup_{\alpha \in [a_x; b_x]} \frac{|L(\alpha) - f_n(\alpha)|}{1 + f_n(\alpha)} \leq \kappa_1 [\log(2 + x) + y + \log n]^2 n^{-\min(\frac{1}{12}, \frac{\delta_4}{2})}.$$

This claim is proved in appendix B.1. We then turn to the approximation of Z. The first step is to approximate Z, conditionally on D_n^T, by a Gaussian process with the same variance-covariance function. To do this, we proceed using strong approximation, as explained in the introduction (equation (1)). This leads to the following claim.

Claim 6. *Let Z be the process given by definition 9. There exists a gaussian process $(Z_\alpha^1)_{\alpha \in \left[\frac{-k_*}{\Delta}; +\infty\right)}$ with the same variance-covariance function as Z: for any $(\alpha_1, \alpha_2) \in [a_x; b_x]^2$, $\mathrm{Cov}(Z_{\alpha_1}^1, Z_{\alpha_2}^1) = \mathrm{Cov}(Z_{\alpha_1}, Z_{\alpha_2})$ and such that for all $n \geq 1$, for all $x > 0$, with probability greater than $1 - e^{-y}$,*

$$\mathbb{E}\left[\sup_{\alpha \in [a_x; b_x]} \frac{|Z_\alpha - Z_\alpha^1|}{1 + f_n(\alpha)} \middle| D_n^T\right] \leq \kappa_5(c_1, \delta_5)(1 + x)(1 + y) n^{-\frac{\delta_5}{3}}.$$

Furthermore, Z^1 can be expressed as $Z^1 = H(D_n^{T^c}, \nu)$, with ν a uniform random variable independent from D_n and H a measurable function from $\mathbb{R}^{T^c} \times [0, 1]$ to $C\left(\left[\frac{-k_*}{\Delta}; +\infty\right), \mathbb{R}\right)$.

This claim is proved in appendix B.2. It implies that the approximation error is negligible up to the scaling factor $\mathfrak{e}$. We will now seek to approximate the process Z^1 given by claim 6 by a time-changed Wiener process. To this end, we first approximate the variance-covariance function of Z^1 (which is the same as that of Z). Let us now give an explicit formula for it.

For all $j \in \mathbb{Z} \cap [a_x \Delta; b_x \Delta]$, by definition 9 of Z:

$$Z\left(\frac{j}{\Delta}\right) = \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right)\left(\hat{s}_{k_*+j}^T - \hat{s}_{k_*}^T\right) = \begin{cases} 0 \text{ if } j = 0, \\ \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right) \sum_{i=k_*+1}^{k_*+j} \hat{\theta}_i^T \psi_i \text{ if } j > 0, \\ \frac{2}{\mathfrak{e}}\left(P_n^{T^c} - P\right) \sum_{i=k_*+j+1}^{k_*} \hat{\theta}_i^T \psi_i \text{ if } j < 0 \end{cases} \tag{19}$$

In other words, for all $j \in \mathbb{Z} \cap [a_x \Delta; b_x \Delta]$,

$$Z\left(\frac{j}{\Delta}\right) = \mathrm{sgn}(j) \frac{2}{\mathfrak{e}} \sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} \hat{\theta}_i^T \left(P_n^{T^c} - P\right)(\psi_i).$$

Let $n_v = |T^c| = n - n_t$. Thus, for any $(j_1, j_2) \in \{a_x \Delta, \ldots, b_x \Delta\}^2$ and any variable X with distribution $s(x)dx$,

$$\begin{aligned} &\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right), Z\left(\tfrac{j_2}{\Delta}\right) | D_n^T\right) \\ &\quad = \mathrm{sgn}(j_1)\mathrm{sgn}(j_2) \frac{4}{n_v \mathfrak{e}^2} \sum_{i_1=k_*-(j_1)_-+1}^{k_*+(j_1)_+} \sum_{i_2=k_*-(j_2)_-+1}^{k_*+(j_2)_+} \hat{\theta}_{i_1}^T \hat{\theta}_{i_2}^T \mathrm{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)). \end{aligned}$$

Let $I = \{k_* - (j_1)_- + 1, \dots, k_* + (j_1)_+\}$ and $J = \{k_* - (j_2)_- + 1, \dots, k_* + (j_2)_+\}$. Assuming concentration around the expectation yields

$$\sum_{i\in I}\sum_{j\in J} \hat{\theta}_i^T \hat{\theta}_j^T \operatorname{Cov}(\psi_i(X), \psi_j(X)) \sim \sum_{i\in I}\sum_{j\in J} \theta_i\theta_j \operatorname{Cov}(\psi_i(X), \psi_j(X)) + \frac{1}{n_t}\sum_{i\in I}\sum_{j\in J} \operatorname{Cov}(\psi_i(X), \psi_j(X))^2. \tag{20}$$

Now, for an increasing function $g : A \to \mathbb{R}$ defined on some interval A containing 0 and such that $g(0) = 0$, the covariance kernel of time-changed symmetric brownian motion $(W_{g(t)})_{t\in A}$ is

$$K(g)(s,t) = \begin{cases} \min(g(s), g(t)) = g(s \wedge t) \text{ if } (s,t) \in (A \cap \mathbb{R}_+)^2 \\ -\max(g(s), g(t)) = -g(s \vee t) \text{ if } (s,t) \in (A \cap \mathbb{R}_-)^2 \\ 0 \text{ else .} \end{cases} \tag{21}$$

Thus, in order to approximate $(s,t) \mapsto \operatorname{Cov}\left(Z(s), Z(t)|D_n^T\right)$ by $K(g)$ for some g, one must show that

$$\begin{aligned} \operatorname{Cov}\left(Z\left(\frac{j_1}{\Delta}\right), Z\left(\frac{j_2}{\Delta}\right)|D_n^T\right) &= \sum_{i\in I}\sum_{j\in J} \hat{\theta}_i^T \hat{\theta}_j^T \operatorname{Cov}(\psi_i(X), \psi_j(X)) \\ &\sim \sum_{i\in I\cap J}\sum_{j\in I\cap J} \theta_i\theta_j \operatorname{Cov}(\psi_i(X), \psi_j(X)) \\ &\quad + \frac{1}{n_t}\sum_{i\in I\cap J}\sum_{j\in I\cap J} \operatorname{Cov}(\psi_i(X), \psi_j(X))^2 \end{aligned} \tag{22}$$

since $I \cap J = \{k_* - (j_1 \vee j_2)_- + 1, \dots, k_* + (j_1 \wedge j_2)_+\}$. This clearly follows from equation (20) if the covariance kernel $(i,j) \mapsto \operatorname{Cov}(\psi_i(X), \psi_j(X))$ is diagonal. Of course, this is not exactly true and the idea is therefore to show that the off-diagonal terms $\operatorname{Cov}(\psi_i(X), \psi_j(X))$ decay fast enough (as $|i-j|$ grows) so that equation (22) still holds approximately.

To see this, remark that for any $(i_1, i_2) \in \mathbb{N}^2$,

$$\psi_{i_1}(X)\psi_{i_2}(X) = 2\cos(2i_1\pi X)\cos(2i_2\pi X) = \cos\left(2(i_1+i_2)\pi X\right) + \cos(2(i_1-i_2)\pi X)$$

and by definition, for all $i \in \mathbb{N}^*$, $\psi_i = \sqrt{2}\cos(2i\pi X)$, while $\psi_0 = 1 = \cos(0\pi x)$. As a result, $\psi_{i_1}(X)\psi_{i_2}(X) = \frac{\psi_{i_1+i_2}(X) + \psi_{|i_1-i_2|}(X)}{\sqrt{2}}$ if $i_1 \neq i_2$ and

$$\operatorname{Cov}(\psi_{i_1}(X), \psi_{i_2}(X)) = \frac{\theta_{i_1+i_2}}{\sqrt{2}} + \left(\frac{1-\delta_{i_1,i_2}}{\sqrt{2}} + \delta_{i_1,i_2}\right)\theta_{|i_2-i_1|} - \theta_{i_1}\theta_{i_2}.$$

By assumption, the sequence $|\theta_k|$ tends to 0 with a polynomial rate of convergence, hence for sequences $i_1 \sim i_2$ tending to $+\infty$, $\theta_{|i_1-i_2|}$ dominates $\theta_{i_1}\theta_{i_2}$ and $\theta_{i_1+i_2}$. Heuristically, it can thus be expected that

$$\sum_{i\in I}\sum_{j\in J} \hat{\theta}_i^T \hat{\theta}_j^T \operatorname{Cov}(\psi_i(X), \psi_j(X)) \sim \sum_{i\in I}\sum_{j\in J} \theta_i\theta_j \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|j-i|} + \sum_{i\in I}\sum_{j\in J} \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)^2 \theta_{|j-i|}^2.$$

The concentration property and the above argument lead to the following proposition, the rigourous proof of which can be found in appendix B.3.

Proposition 6.1. *Let P be the probability measure with pdf s on $[0;1]$, let $\theta_j = \langle s, \psi_j \rangle = P(\psi_j)$ and assume that the coefficients θ_j satisfy the hypotheses of section 3. For any integer interval I, let*

$$w(I) = 1 + \frac{|I|}{\Delta} + \frac{1}{\mathcal{E}} \sum_{j \in I} \theta_j^2.$$

Let $\hat{\theta}_j^T = P^T(\psi_j)$ and let $I^1, I^2 \subset \{k_* + a_x\Delta, \dots, k_* + b_x\Delta\}$ be two intervals. Then the statistics

$$U_{I^1,I^2} = \sum_{i \in I^1} \sum_{j \in I^2} \hat{\theta}_i^T \hat{\theta}_j^T [P(\psi_i \psi_j) - P\psi_i P\psi_j]$$

can be approximated in the following way: there exists two constants κ_4 and $u_3 > 0$ such that, with probability greater than $1 - e^{-y}$,

$$\begin{aligned} U_{I^1,I^2} = \frac{1}{2}\frac{|I^1 \cap I_k^2|}{n_t} + \left(1 - \frac{1}{\sqrt{2}}\right) \sum_{i \in I^1 \cap I^2} \theta_i^2 + \frac{1}{\sqrt{2}} \sum_{i \in I^1} \sum_{j \in I^2} \theta_i \theta_j \theta_{|i-j|} + \frac{1}{2n_t} \sum_{i \in I^1} \sum_{j \in I^2} \theta_{|i-j|}^2 \\ \pm \kappa_4 (y + \log n)^2 (1+x)^{\frac{3}{2}} \sqrt{w(I_1) w(I_2)} n^{-u_3} \mathcal{E}. \end{aligned}$$

Now , the decay of the θ_k afforded by hypothesis 1 must be exploited in order to show that

$$U_{I,J} \approx U_{I \cap J, I \cap J}$$

or equivalently, that

$$U_{I,J \setminus I} + U_{I \setminus J, I} = U_{I,J} - U_{I \cap J, I \cap J}$$

is negligible. That is the point of the following claim.

Claim 7. *Let*

$$(k_1, k_2) \in \{k_* + a_x\Delta, \dots, k_* + b_x\Delta\}^2$$

and let $m_1 < m_2 < m_3$ be the increasing re-ordering of $\{k_*, k_1, k_2\}$ (assuming it exists). For any integer $j \geq -k_*$, let

$$w_0(j) = 1 + \frac{|j|}{\Delta} + \frac{1}{\mathcal{E}} \sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} \theta_i^2.$$

Under the assumptions of Theorem 1, there exists constants $\kappa_7 \geq 0, u_2 > 0$ such that

$$\sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{j_1} \theta_{j_2} \theta_{|j_1-j_2|} \leq \kappa_7 \sqrt{w_0(k_1 - k_*) w_0(k_2 - k_*)} \left[(1+x)^{\frac{\delta_6}{3\delta_2}} \mathcal{E}^{\frac{2}{3}} \left(\frac{k_*^{\delta_3 - \delta_6}}{n_t} \right)^{\frac{1}{3}} + n^{-\frac{2u_2}{3}} \mathcal{E} \right] \tag{23}$$

$$\frac{1}{n_t} \sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{|j_1-j_2|}^2 \leq \kappa_7 \sqrt{w_0(k_1 - k_*) w_0(k_2 - k_*)} n^{-\frac{2\delta_4}{3}} \mathcal{E}, \tag{24}$$

Moreover, let

$$E_{m_1,m_2,m_3} = \sum_{j_1=m_{(1)}+1}^{m_{(2)}} \sum_{j_2=m_{(2)}+1}^{m_{(3)}} \hat{\theta}_{j_1}^T \hat{\theta}_{j_2}^T \operatorname{Cov}(\psi_{j_1}(X), \operatorname{Cov}(\psi_{j_2}(X))$$

It follows that with probability greater than $1-e^{-y}$,

$$|E_{k_*,k_1,k_2}| \leq \kappa_7(1+x)^{u_6}(y+\log n)^2\sqrt{w_0(k_1-k_*)w_0(k_2-k_*)}\left(n^{-u_4}\mathcal{E}+\mathcal{E}^{\frac{2}{3}}\left(\frac{k_*^{\delta_3-\delta_6}}{n_t}\right)^{\frac{1}{3}}\right), \tag{25}$$

where $u_4>0, u_6$ *are constants.*

This claim is proved in appendix B.4. To approximate the covariance kernel by $K(g_n)$, it remains to prove that $\mathrm{Var}\left(Z(\alpha)|D_n^T\right)$ can be expressed (approximately) as an increasing function $g_n(\alpha)$, where g_n satisfies points 1-3 of theorem 1. This leads to the following claim, which is proved in section 2 of the supplementary material.

Claim 8. *There exists a function* g_n *satisfying the conditions of Theorem 1 and a constant* $u_5>0$ *such that, for all* $x,y>0$*, with probability greater than* $1-e^{-y}$*, for all* $(j_1,j_2)\in[a_x\Delta,\dots,b_x\Delta]$,

$$\left|\mathrm{Cov}\left(Z\left(\tfrac{j_1}{\Delta}\right),Z\left(\tfrac{j_2}{\Delta}\right)|D_n^T\right)-K(g_n)\left(\tfrac{j_1}{\Delta},\tfrac{j_2}{\Delta}\right)\right|\leq\kappa_6(1+x)^{u_6}[y+\log n]^2\sqrt{w_0(j_1)w_0(j_2)}\left(n^{-u_5}+\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\right)^{\frac{1}{3}}\right),$$

where $\kappa_6\geq 0, u_5>0, u_6\geq 0$ *are constants and*

$$w_0(j)=1+\frac{|j|}{\Delta}+\frac{1}{\mathcal{E}}\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+}\theta_i^2$$

for all $j\in\mathbb{Z}\cap[-k_*;+\infty)$

Thus, the conditional variance-covariance function of the process Z^1 is uniformly close to that of the process W_{g_n}. To conclude the proof of Theorem 1, it remains to derive a uniformly close coupling of these two Gaussian processes from the above results on their covariance kernels. This is accomplished using the following proposition, the proof of which can be found in section 5 of the supplementary material.

Proposition 6.2. *Let* $([x_i,x_{i+1}))_{1\leq i\leq M-1}$ *be a partition of the interval* $[a,b)$*. Let* $Y:\{x_1,\dots,x_M\}\to\mathbb{R}$ *be such that* $(Y(x_j))_{1\leq j\leq M}$ *is a zero-mean gaussian vector. Abusing notation, we also denote by* Y *the extension of* Y *to* $[a;b]$ *by linear interpolation. Let* $K_Y:[a;b]^2\to\mathbb{R}$ *be the variance-covariance function of* Y*. Let* $h:[a;b]\to\mathbb{R}$ *be a continuous, increasing functions and let* $K_X:[a;b]^2\to\mathbb{R}$ *be a positive semi-definite function such that:*

$$\forall(s,t)\in[a;b]^2,\ |K_X(s,s)+K_X(t,t)-2K_X(s,t)|\leq|h(s)-h(t)|.$$

Assume that there exists constants $L>0$ *and* $\varepsilon\in[0;1]$ *such that:*

- $\sup_{t\in[a;b]}\sqrt{\frac{K_X(t,t)}{1+|h(t)|}}\leq L$
- *For any* $i\in\{1,\dots,M-1\}, |h(x_{i+1})-h(x_i)|\leq\varepsilon(1+|h(x_{i+1})|\vee|h(x_i)|)$
- *For all* $(i,j)\in\{1,\dots,M\}^2$, $|K_X(x_i,x_j)-K_Y(x_i,x_j)|\leq\varepsilon\sqrt{(1+|h(x_i)|)(1+|h(x_j)|)}$.

There exists a universal constant κ *and a measurable function* $f:C([a;b],\mathbb{R})\to C([a;b],\mathbb{R})$ *such that for all random variables* $\nu\sim\mathcal{U}([0;1])$ *independent from* Y*,* $X=f(Y,\nu)$ *is a zero-mean gaussian process with variance-covariance function* K_X *and moreover,*

$$\mathbb{E}\left[\sup_{a\leq t\leq b}\frac{|X_t-Y_t|}{\sqrt{1+|h(t)|}}\right]\leq\kappa\sqrt{L\log(M+1+\max(|h(b)|,|h(a)|))}\varepsilon^{\frac{1}{12}}.$$

The proof of this proposition uses the known optimal Wasserstein coupling between (finite-dimensional) random vectors, together with the Kolmogorov continuity theorem to pass to the continuum limit.

The proof of Theorem 1 then concludes as follows. Let R be as in Definition 7. Let

$$h_0 : \alpha \mapsto \frac{1}{\mathcal{E}} \left(R(k_*) - R(k_* + \alpha\Delta) \right)$$

and remark that, by Definition 7, for all $\alpha \in [a_x; b_x]$,

$$h_0(\alpha) = \alpha - \frac{\mathfrak{e}}{\mathcal{E}} f_n(\alpha) \tag{26}$$

Let then

$$h : \alpha \mapsto 4 \left\| s \right\|^2 \alpha + 8 \left\| s \right\|_\infty h_0(\alpha). \tag{27}$$

By claim 8 above,

$$|K(g_n)(\alpha, \alpha) + K(g_n)(\alpha', \alpha') - 2K(g_n)(\alpha', \alpha)| = |g_n(\alpha) - g_n(\alpha')| \leq |h(\alpha) - h(\alpha')|$$

for any α, α'. We now apply Proposition 6.2 to $K_X = K(g_n)$ and $Y = Z$, on the interval $[a_x; b_x]$. Z is piecewise linear between the points $x_i = \frac{i}{\Delta}$, which number at most $M = (b_x - a_x)\Delta$.

Let us now prove the three bullet points. First, since $g_n(0) = 0$, for any $\alpha \in [a_x; b_x]$,

$$K(g_n)(\alpha, \alpha) = |g_n(\alpha)| \leq |h(\alpha)|.$$

Secondly, by definition of h, h_0, for any $j \in [a_x\Delta + 1; b_x\Delta]$,

$$\begin{aligned} \left| h\left(\frac{j}{\Delta}\right) - h\left(\frac{j-1}{\Delta}\right) \right| &= \frac{4 \left\| s \right\|^2}{\Delta} + 8 \left\| s \right\|_\infty \left| h_0\left(\frac{j}{\Delta}\right) - h_0\left(\frac{j-1}{\Delta}\right) \right| \\ &\leq \frac{4 \left\| s \right\|^2}{\Delta} + \frac{8 \left\| s \right\|_\infty}{\mathcal{E}} \max\left(\theta_j^2, \theta_{j-1}^2\right) \end{aligned}$$

By claim 9 in the appendix, since $\Delta \geq \frac{n_t}{n - n_t} \geq n^{\delta_4}$ by lemma 5.3 and hypothesis 5,

$$\begin{aligned} \left| h\left(\frac{j}{\Delta}\right) - h\left(\frac{j-1}{\Delta}\right) \right| &\leq 4 \left\| s \right\|^2 n^{-\delta_4} + 8 \left\| s \right\|_\infty \left(c_3 c_6 \frac{k_*^{\delta_3 - \delta_6}}{n_t \mathcal{E}} + \kappa (1+x)^{\frac{\delta_6}{\delta_2}} n^{-u_2} \left| h\left(\frac{j-1}{\Delta}\right) \right| \vee \left| h\left(\frac{j}{\Delta}\right) \right| \right) \\ &\leq \left(4 \left\| s \right\|^2 + 8 \left\| s \right\|_\infty \right) \left(c_3 c_6 \frac{k_*^{\delta_3 - \delta_6}}{\Delta} + n^{-u_2 \wedge \delta_4} \right) \left[1 + \left| h\left(\frac{j-1}{\Delta}\right) \right| \vee \left| h\left(\frac{j}{\Delta}\right) \right| \right] \end{aligned}$$

which proves the condition in the second bullet point, for any

$$\varepsilon \geq \left(4 \left\| s \right\|^2 + 8 \left\| s \right\|_\infty \right) \left(c_3 c_6 \frac{k_*^{\delta_3 - \delta_6}}{\Delta} + n^{-u_2 \wedge \delta_4} \right).$$

Finally, for any $\alpha = \frac{j}{\Delta}$,

$$\begin{aligned} w_0(\alpha\Delta) &= 1 + \frac{|j|}{\Delta} + \frac{1}{\mathcal{E}} \sum_{i = k_* - (j)_- + 1}^{k_* + (j)_+} \theta_i^2 \\ &= 1 + |\alpha| + |h_0(\alpha)| \\ &\leq 1 + |h(\alpha)| \end{aligned}$$

since $\|s\|^2 \geq 1$ and $\|s\|_\infty \geq 1$. By Claim 8, the condition in the third bullet point holds for any

$$\varepsilon \geq \kappa_6(1+x)^{u_6}[y+\log n]^2\left(n^{-u_5}+\left(\frac{k_*^{\delta_3-\delta_6}}{\Delta}\right)^{\frac{1}{3}}\right)$$

whenever $D_n^T \in E_y$, where E_y is an event with probability greater than $1-e^{-y}$. Thus, Proposition 6.2 applies to the process Z_1 conditionally on $D_n^T \in E_y$ with the function h given by equation (27). To conclude, it remains to see that $\sqrt{1+|h|} \leq \kappa(1+f_n)$ for some constant κ depending on s only. This follows from the fact that $h_0(\alpha) = \alpha - \frac{\mathfrak{e}}{\mathcal{E}} f_n(\alpha)$ (equation (26)), which yields

$$\begin{aligned}
1+|h(\alpha)| &\leq (4\|s\|^2 \vee 8\|s\|_\infty)(1+|\alpha|+|h_0(\alpha)|) \\
&\leq 8\|s\|_\infty (1+2|\alpha|+f_n(\alpha)) \text{ by lemma 5.3} \\
&\leq 8\|s\|_\infty (3+2f_n^2(\alpha)+f_n(\alpha)) \text{ by lemma 5.1} \\
&\leq 24\|s\|_\infty (1+f_n(\alpha))^2. \qquad (28)
\end{aligned}$$

The remaining technical details are handled in appendix B.5.

7 Discussion

In this section, we interpret the results of the article and discuss how they could be extended.

7.1 Implications of our results

Theorem 1 provides an approximation to the rescaled hold-out process, with scale factor Δ given by Definition 5. The approximating process is asymptotically random, tight along a sequence of intervals of length greater than 1 and trends away from zero at $\pm\infty$ (as discussed in section 5.1). In the case of "incomplete" cross-validation for fixed V, the process is of identical type, but with the variance reduced by a factor V. This means that CV performs significantly better than the hold-out when comparing $\hat{s}_k^T$ and $\hat{s}_{k_*}^T$ for k close to $k_*(n_t)$. If $V \to +\infty$, then rescaled cross-validation concentrates around the deterministic function f_n, which suggests that CV has asymptotically better performance than the hold-out. This supports the belief that cross-validation improves when V is increased. Though $k_*(n_t)$ is not the oracle based on the full sample $(k_*(n))$, the greater concentration of CV around $k_*(n_t)$ makes it possible to choose n_t closer to n than would be reasonable for the hold-out, resulting in improved overall performance.

7.2 Nested models

For technical convenience, the hold-out was analysed for a specific collection of estimators in least-squares density estimation. We now consider possible applications of the methods and results of this article to other model selection problems.

In order to obtain a Brownian approximation similar to that of Theorem 1, the estimators should be linearly ordered, so that there is, in effect, a single real hyperparameter. Moreover, to carry out computations similar to ours, these estimators should be least-squares estimators based on linear models. Thirdly, the martingale property of the limit seems linked to the fact that models are *nested*, which allows to write the estimators as nested partial sums. Finally, our main hypotheses require the risk to decrease *polynomially* with

respect to the parameter. In the trigonometric case, this is justified through the connection with regularity (differentiability) of the target function, and this connection should hold also for other model collections.

Let us therefore consider a collection of nested, linear models

$$m_1 \subset m_2 \subset \ldots \subset m_k \subset \ldots$$

and their corresponding least-squares estimators $(\hat{s}_k)_{k\geq 1}$. There are many examples of such collections, but we have in mind mainly the following:

- Naturally ordered $1D$ orthogonal bases, such as the Legendre polynomials, Hermite polynomials, etc.
- The spherical harmonics, ordered by degree. More generally, the eigenfunctions of the Laplace operator on a manifold, ordered by eigenvalue.
- Multivariate Fourier series, using either square partial sums - corresponding for example to

  $$m_k = \left\{1, \sqrt{2}\cos(2\pi lx), \sqrt{2}\cos(2\pi my), 2\cos(2\pi lx)\cos(2\pi my) : 1 \leq l \leq k, 1 \leq m \leq k\right\},$$

  or circular partial sums, corresponding to

  $$\left\{1, \sqrt{2}\cos(2\pi lx), \sqrt{2}\cos(2\pi my), 2\cos(2\pi lx)\cos(2\pi my) : l^2 + m^2 \leq k\right\}.$$

Let us now briefly outline the main ingredients of the analysis in the general case.

- Centering and scaling: our definitions used the fact that $\mathrm{Var}_P(\psi_j) \sim 1$, for the trigonometric basis functions ψ_j. For a general basis, the definitions should be adapted to take into account a variance which may depend on j.
- Concentration of the risk (claim 5): there should be little difficulty in generalizing this part, since the key result used [2, Lemma 14] applies to any linear model.
- Gaussian approximation (claim 6): for univariate density estimation, the same argument may be used, provided only that $\|\hat{s}_k\|_{L^1}$ can be suitably bounded. For multivariate models, the method must be modified. For example, one may use iterated integration by parts to write

  $$(P_n - P)(f) = (-1)^d \int_{\mathbb{R}^d} \partial_1\partial_2..\partial_d f(x_1, \ldots, x_d)(\hat{F}_n - F)(x_1, \ldots, x_d)dx_1..dx_d,$$

  where F is the multivariate distribution function and $\hat{F}_n$ its empirical counterpart. One may then appeal to a strong approximation theorem for the multivariate empirical process, as can be found for example in [9].
- Approximation of the covariance kernel: many computations remain valid for a general basis $(\psi_i)_{i\geq 1}$, covariance kernel $c_{i,j} = \mathrm{Cov}_P(\psi_i, \psi_j)$ and coefficients $\theta_i = P\psi_i$. Of those that rely on the specific form of the covariance kernel, the main property used is that

  $$c_{i,j} \approx \theta_{|i-j|} \text{ where } \sum_{k\geq r} |\theta_k| = o\left(r^{-\delta}\right) \text{ by assumption.}$$

  Qualitatively, this means that the covariance is small off of the diagonal. A similar assumption can likely be made for other models, though its exact formulation may vary.

7.3 Hard thresholding and wavelets

Let us briefly discuss *hard thresholding*, a method often used in conjunction with the wavelet basis. When used together with model selection, the effect of *hard thresholding* is to re-order the initial basis $(\hat{\psi}_j)_{j\geq 1}$ in a data-dependent way as $(\psi_{\hat{j}})_{j\geq 1}$, where $\hat{1}, \hat{2}, ...$ is a permutation of $\mathbb{N}$ which renders the sequence

$$|\hat{\theta}^T_{\hat{j}}| = \left| \frac{1}{|T|} \sum_{i\in T} \psi_{\hat{j}}(X_i) \right|$$

non-increasing. Since the basis is then ordered, it makes sense to look for a one-dimensional approximating process.

The corresponding estimators

$$\hat{s}^T_k = \sum_{j=1}^{k} \hat{\theta}^T_{\hat{j}} \psi_{\hat{j}},$$

are non-linear because of the data-dependent basis. Part of our analysis still applies to this method, since conditioning on the training data fixes the basis. In particular, the Gaussian approximation (claim 6) should still work. However, questions regarding concentration of the risk or of the conditional covariance seem much harder.

7.4 Model selection

The CV risk estimator is usually used to select a model, $\hat{k}^{cv}_{\mathcal{T}}$, which minimizes it. The final result of CV is then the estimator $\hat{s}_{\hat{k}^{cv}_{\mathcal{T}}}$, or $\hat{s}^T_{\hat{k}_T}$ in the case of simple validation. Thus, what we are most interested in practice is the risk $\left\| \hat{s}_{\hat{k}^{cv}_{\mathcal{T}}} - s \right\|^2$ of this final estimator, and how it depends on the CV method used (at least when CV is used with a goal of *estimation*, as opposed to *identification* of the best model). There are several ways our results can contribute to answering these questions. First, it follows from proposition 5.2 and Markov's inequality that $\frac{\hat{k}^{cv}_{\mathcal{T}}}{\Delta} \in [a_x, b_x]$ with high probability for x large enough. Since $\hat{R}^{cv}_{\mathcal{T}}(u)$ can be uniformly approximated by $f_n(u) - W_{g_n(u)/V}$ on $[a_x, b_x]$, it is natural to approximate $\frac{\hat{k}^{cv}_{\mathcal{T}} - k_*}{\Delta}$ (the minimizer of $\hat{R}^{cv}_{\mathcal{T}}(u)$) by the minimizer $\hat{\alpha}_{n,V}$ of $f_n - W_{g_n/V}$. The minima of $f_n - W_{g_n/V}$ can be studied using the theory of Wiener processes. Together with our results about f_n and g_n (in lemma 5.1 and Theorem 1), this makes the analysis of $\hat{\alpha}_{n,V}$ much easier than that of $\hat{k}^{cv}_{\mathcal{T}}$. Moreover, claim 5 of this article proves that the (excess) risk $\left\| \hat{s}^T_k - s \right\|^2 - \left\| \hat{s}^T_{k_*} - s \right\|^2$ concentrates around $\mathfrak{e} f_n\left(\frac{k-k_*}{\Delta}\right)$ for k "close enough" to k_*. This removes the dependency on the training sample D^T_n and reduces the analysis of $\left\| \hat{s}^T_k - s \right\|^2 - \left\| \hat{s}^T_{k_*} - s \right\|^2$ to that of the deterministic function $f_n\left(\frac{k-k_*}{\Delta}\right)$.

7.5 Other cross-validation methods

In this article, we only considered a particular type of cross-validation. Corollary 2 relies on decomposing "incomplete" $V-$fold CV as a finite average of asymptotically independent hold-out estimators, which can be analysed more easily than general cross-validation because conditionally on their training data, they are empirical processes.

In general, for projection estimators in L^2 density estimation, the cross-validation risk estimator is not a (conditional) empirical process but a (weighted) U-statistic of order 2 (more precisely, a weighted sum of the

terms $\psi_k(X_i)\psi_k(X_j)$). Thus, new methods are required to approximate general CV estimators by gaussian processes. We conjecture that more general cross-validation methods also behave locally like the sum of the (rescaled) excess risk and a time-changed Wiener process, though we expect the scaling and the time-change g_n to be different for different versions of CV.

Theorem 1 can also shed light on the behaviour of other methods which use simple validation as a key ingredient, such as Aggregated hold-out [17].

A Preliminary results

In all appendices, the term constant means a function of $\|s\|_\infty, \|s\|^2$ and the constants $c_1, c_2, c_3, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6$, which appear in the hypotheses of Theorem 1. Note that by hypothesis (1), $\|\theta\|_{\ell^1}, \|s\|_\infty, \|s\|^2$ are finite and can be bounded by functions of c_1, δ_1. The letter u will denote strictly positive constants that only depend on $(\delta_i)_{1\leq i\leq 6}$ (they will generally appear as exponents of $\frac{1}{n}$). The letter κ denotes a non-negative constant. The notation $n_v = n - n_t$ will also be used frequently. The results of this section are independent from the rest. They will be used in the rest of the proof of Theorem 1.

A.1 Proof of lemma 5.3

- By definition and non-negativity of θ_j^2, $\sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}} \leq 1$, therefore $\Delta \geq \Delta_d \geq \frac{n_t}{n-n_t}$.
- $\mathcal{E} = \frac{\Delta}{n_t} \geq \frac{1}{n-n_t}$.
- $\mathfrak{e} = \sqrt{\frac{\mathcal{E}}{n-n_t}} \geq \sqrt{\frac{1}{(n-n_t)^2}} = \frac{1}{n-n_t}$.
- $\frac{\mathcal{E}}{\mathfrak{e}} = \mathcal{E}\sqrt{\frac{n-n_t}{\mathcal{E}}} = \sqrt{(n-n_t)\mathcal{E}} \geq 1$.
- By definition, $\Delta_g \leq k_*$. Thus $\frac{\Delta_g}{n_t} \leq \frac{k_*}{n_t} \leq \mathrm{or}(n_t)$. Moreover,

$$\left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}}\right]\frac{1}{n_t} \leq \frac{1}{\Delta_d}\sum_{j=k_*+1}^{k_*+\Delta_d}\theta_j^2 \leq \frac{1}{\Delta_d}\sum_{j=k_*+1}^{+\infty}\theta_j^2 \leq \frac{\mathrm{or}(n_t)}{\Delta_d}.$$

Thus

$$\begin{aligned} n_t\mathrm{or}(n_t) &\geq \Delta_d - \sqrt{\frac{n_t}{n-n_t}}\sqrt{\Delta_d} \\ &\geq \Delta_d - \frac{1}{2}\frac{n_t}{n-n_t} - \frac{1}{2}\Delta_d \\ &\geq \frac{1}{2}\Delta_d - \frac{1}{2}\frac{n_t}{n-n_t}. \end{aligned}$$

It follows that

$$\Delta_d \leq 2n_t\mathrm{or}(n_t) + \frac{n_t}{n-n_t},$$

so since $\frac{n_t}{n-n_t}\frac{1}{n_t} = \frac{1}{n-n_t}$,

$$\frac{\Delta_d}{n_t} \leq 2\mathrm{or}(n_t) + \frac{1}{n-n_t},$$

which proves the result.

A.2 Proof of lemma 5.1

By definition,

$$\sum_{j=1}^{\Delta_d} \theta^2_{k_*+j} \geq \frac{\Delta_d}{n_t}\left(1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_d}}\right) = \frac{\Delta_d}{n_t} - \sqrt{\frac{\Delta_d}{n_t(n-n_t)}}$$

$$\begin{aligned}\sum_{j=0}^{\Delta_g - 1} \theta^2_{k_*-j} &\leq \frac{\Delta_g}{n_t}\left(1 + \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta_g}}\right) \\ &\leq \frac{\Delta_g}{n_t} + \sqrt{\frac{\Delta_g}{n_t(n-n_t)}}.\end{aligned}$$

It follows that

$$f_n\left(\frac{\Delta_d}{\Delta}\right) \leq \frac{1}{\mathfrak{e}}\sqrt{\frac{\Delta_d}{n_t(n-n_t)}}$$

$$f_n\left(\frac{-\Delta_g}{\Delta}\right) \leq \frac{1}{\mathfrak{e}}\sqrt{\frac{\Delta_g}{n_t(n-n_t)}}$$

Choosing $x = 1$ if $\Delta = \Delta_d$ and $x = -1$ otherwise yields the first bound.

Consider now the second bound. It is trivially valid when $|\alpha| \leq 1$, since $f_n(\alpha) \geq 0$. Let now α be such that $|\alpha| > 1$. Consider first the case $\alpha = \frac{l}{\Delta}$ for some $l \in \mathbb{Z}$ such that $|l| > \Delta$. If $l > \Delta \geq \Delta_d$, then by definition,

$$\sum_{j=1}^{l} \theta^2_{k_*+j} < \frac{l}{n_t}\left(1 - \sqrt{\frac{n_t}{l(n-n_t)}}\right) \leq \frac{l}{n_t} - \sqrt{\frac{l}{n_t(n-n_t)}}$$

which implies that

$$f_n\left(\frac{l}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left[\frac{l}{n_t} - \sum_{j=1}^{l}\theta^2_{k_*+j}\right] > \frac{1}{\mathfrak{e}}\sqrt{\frac{l}{n_t(n-n_t)}} = \sqrt{\frac{l}{\Delta}}.$$

If now $l < -\Delta \leq -\Delta_g$, then by definition of Δ_g,

$$\sum_{j=l+1}^{0} \theta^2_{k_*+j} > \frac{|l|}{n_t}\left(1 + \sqrt{\frac{n_t}{|l|(n-n_t)}}\right) = \frac{|l|}{n_t} + \sqrt{\frac{|l|}{n_t(n-n_t)}}$$

which implies that

$$f_n\left(\frac{l}{\Delta}\right) = \frac{1}{\mathfrak{e}}\left[\sum_{j=l+1}^{0}\theta^2_{k_*+j} - \frac{|l|}{n_t}\right] > \frac{1}{\mathfrak{e}}\sqrt{\frac{|l|}{n_t(n-n_t)}} = \sqrt{\frac{|l|}{\Delta}}.$$

Let now $\alpha \in \mathbb{R}$ such that $|\alpha| > 1$. Let $l \in \mathbb{Z}$ and $\theta \in (0;1]$ be such that

$$\alpha = \theta\frac{l}{\Delta} + (1-\theta)\frac{l-1}{\Delta}$$

if $\alpha > 0$ and

$$\alpha = \theta \frac{l}{\Delta} + (1-\theta)\frac{l+1}{\Delta}$$

if $\alpha < 0$. By piecewise linearity of f_n,

$$f_n(\alpha) = \theta f_n\left(\frac{l}{\Delta}\right) + (1-\theta) f_n\left(\frac{l-\operatorname{sgn}(\alpha)}{\Delta}\right).$$

We necessarily have that $|l| > |l| - 1 \geq \Delta$. If $|l| = \Delta + 1$, then

$$f_n(\alpha) \geq \theta\sqrt{\frac{|l|}{\Delta}} = \theta\sqrt{1+\frac{1}{\Delta}} \geq \frac{\theta}{2\Delta} \geq \sqrt{1+\frac{\theta}{\Delta}} - 1 = \sqrt{|\alpha|} - 1.$$

Otherwise, $|l| - 1 \geq \Delta$, hence

$$\begin{aligned}
\sqrt{|\alpha|} &= \sqrt{\frac{|l|-1}{\Delta} + \frac{1-\theta}{\Delta}} \\
&\leq \sqrt{\frac{|l|-1}{\Delta}} + \frac{1-\theta}{2\Delta} \\
&\leq \theta\sqrt{\frac{|l|}{\Delta}} + (1-\theta)\sqrt{\frac{|l|-1}{\Delta}} + \frac{1-\theta}{2\Delta} \\
&\leq f_n(\alpha) + \frac{1-\theta}{2\Delta}.
\end{aligned}$$

This yields the result since $\Delta \geq 1$.

A.3 Some more bounds involving f_n

Lemme A.1. *Let $\alpha \in \mathbb{R}$ be as defined in Theorem1. Then for all $x > 0$,*

$$\frac{|\alpha|\Delta}{n_t} \leq (1 + f_n(\alpha)) \left[2or(n_t) + \frac{1}{n - n_t}\right] \tag{29}$$

$$\frac{|\alpha|\Delta}{n_t} \leq \mathfrak{e}\,(1 + f_n(\alpha)) \sqrt{2(n-n_t)or(n_t) + 2} \tag{30}$$

Moreover,

$$\left|\sum_{k_* - (\alpha\Delta)_- + 1}^{k_* + (\alpha\Delta)_+} \theta_j^2\right| \leq |\alpha|\mathcal{E} + f_n(\alpha)\mathfrak{e} \tag{31}$$

$$\leq 2(1 + f_n(\alpha))^2 \mathcal{E}. \tag{32}$$

Proof. Define, for any $k \geq 1$,

$$l_d(k) = \max\left\{l \in \mathbb{N} : \frac{1}{l}\sum_{j=1}^{l} \theta_{k_*+j}^2 \geq \left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{k}}\right]\frac{1}{n_t}\right\}$$

$$l_g(k) = \max\left\{l \in \{0, \ldots, k_*\} : \frac{1}{l}\sum_{j=0}^{l-1} \theta_{k_*+j}^2 \leq \left[1 + \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{k}}\right]\frac{1}{n_t}\right\}.$$

Define then the three following quantities

$$\begin{aligned}\Delta'_d &= \max\{k \in \mathbb{N} : k \leq l_d(k)\}\\ \Delta'_g &= \max\{k \in \mathbb{N} : k \leq l_g(k)\}\\ \Delta' &= \max(\Delta'_g, \Delta'_d).\end{aligned}$$

Note the similarity with the definition of $\Delta_d, \Delta_g, \Delta$: indeed, if θ_j^2 is a non-increasing sequence, it is easy to see that these quantities coincide. One can further see that, in general, $\Delta'_d \geq \Delta_d$ and $\Delta'_g \geq \Delta_g$. As for upper bounds, it true in general that

$$\Delta'_g \leq l_g(\Delta'_g) \leq k_* \leq n_t or(n_t)$$

and that

$$\begin{aligned}or(n_t) &\geq \sum_{j=1}^{l_d(\Delta'_d)} \theta^2_{k_*+j}\\ &\geq \left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta'_d}}\right]\frac{l_d(\Delta'_d)}{n_t}\\ &\geq \left[1 - \sqrt{\frac{n_t}{n-n_t}}\frac{1}{\sqrt{\Delta'_d}}\right]\frac{\Delta'_d}{n_t}\\ &= \frac{\Delta'_d}{n_t} - \sqrt{\frac{\Delta'_d}{n_t(n-n_t)}}\\ &\geq \frac{\Delta'_d}{2n_t} - \frac{1}{2(n-n_t)}.\end{aligned}$$

This yields

$$\frac{\Delta'}{n_t} \leq 2or(n_t) + \frac{1}{n-n_t}. \tag{33}$$

Let now $\alpha \in \frac{1}{\Delta}\mathbb{Z}$. On the one hand, if $|\alpha|\Delta \leq \Delta'$, then

$$|\alpha|\frac{\Delta}{n_t} \leq \frac{\Delta'}{n_t} \leq 2or(n_t) + \frac{1}{n-n_t}$$

and moreover, by lemma 5.1,

$$\begin{aligned}\frac{|\alpha|\Delta}{n_t} &= |\alpha|\mathcal{E}\\ &= \sqrt{|\alpha|}\sqrt{\frac{\mathcal{E}}{n_v}}\sqrt{n_v|\alpha|\mathcal{E}}\\ &\leq (1+f_n(\alpha))\mathfrak{e}\sqrt{n_v\frac{\Delta'}{n_t}}\\ &\leq (1+f_n(\alpha))\mathfrak{e}\sqrt{2n_v or(n_t)+1}.\end{aligned}$$

On the other hand, if $|\alpha|\Delta > \Delta'$ then either $\alpha\Delta > \Delta'_d$ or $-\alpha\Delta > \Delta'_g$. If $l = \alpha\Delta > \Delta'_d$, then by definition,

$$l \geq \Delta'_d + 1 > l_d\left(\Delta'_d + 1\right),$$

which implies that

$$\frac{1}{l}\sum_{j=k_*+1}^{k_*+l}\theta_j^2 < \left(1-\sqrt{\frac{n_t}{(n-n_t)(\Delta'_d+1)}}\right)\frac{1}{n_t}$$

and hence that

$$\mathfrak{e} f_n(\alpha) = \frac{l}{n_t} - \sum_{j=k_*+1}^{k_*+l}\theta_j^2 \geq \frac{l}{\sqrt{n_t(n-n_t)(\Delta'_d+1)}} = \alpha\sqrt{\frac{\Delta}{\Delta'_d+1}}\mathfrak{e}.$$

Thus,

$$\alpha \leq f_n(\alpha)\sqrt{\frac{\Delta'_d+1}{\Delta}} \leq f_n(\alpha)\sqrt{\frac{\Delta'+1}{\Delta}}.$$

A similar argument yields that

$$|\alpha| \leq f_n(\alpha)\sqrt{\frac{\Delta'_g+1}{\Delta}} \leq f_n(\alpha)\sqrt{\frac{\Delta'+1}{\Delta}}$$

whenever $\alpha\Delta < -\Delta'$. It follows that whenever $|\alpha|\Delta > \Delta'$,

$$|\alpha|\frac{\Delta}{n_t} \leq f_n(\alpha)\frac{\sqrt{\Delta(\Delta'+1)}}{n_t} \leq f_n(\alpha)\frac{\Delta'+1}{n_t} \leq f_n(\alpha)\left[2or(n_t)+\frac{1}{n-n_t}+\frac{1}{n_t}\right]$$

and moreover

$$\begin{aligned}\frac{|\alpha|\Delta}{n_t} &\leq f_n(\alpha)\frac{\sqrt{\Delta(\Delta'+1)}}{n_t}\\ &= f_n(\alpha)\mathfrak{e}\sqrt{n_v\frac{\Delta'+1}{n_t}}\\ &\leq f_n(\alpha)\mathfrak{e}\sqrt{2n_v or(n_t)+2}\end{aligned}$$

using equation (33) and the assumption that $n_t \geq \frac{n}{2}$. Thus, the first bounds hold for all $\alpha \in \frac{1}{\Delta}\mathbb{Z}$, and thus for all $\alpha \in \mathbb{R}$ by piecewise linearity of the functions involved.

Let $l = \alpha\Delta \in \mathbb{Z}$. The second bound follows from the decomposition

$$\sum_{j=k_*-(l)_-+1}^{k_*+(l)_+}\theta_j^2 = \sum_{j=k_*-(l)_-+1}^{k_*+(l)_+}\left[\theta_j^2-\frac{1}{n_t}\right]+\frac{|l|}{n_t} \leq f_n(\alpha)\mathfrak{e}+\frac{|\alpha|\Delta}{n_t} = f_n(\alpha)\mathfrak{e}+|\alpha|\mathcal{E}.$$

□

B Proof of Theorem 1

B.1 Proof of claim 5

Let $j \in \{a_x\Delta,\dots,b_x\Delta\}$. Since $\hat{s}_k^T = \sum_{j=1}^k P_n^T(\psi_j)\psi_j = \sum_{j=1}^k \hat{\theta}_j^T\psi_j$,

$$\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 = \mathrm{sgn}(j)\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+}\left(\hat{\theta}_i^T-\theta_i\right)^2-\theta_i^2.$$

It has been proved [2, Lemma 14] based on a more general result of Lerasle [16, Proposition 6.3] that processes like

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 = \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} (P^T - P)(\psi_j)^2$$

concentrate around their expectation, so that

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 \sim \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \frac{\mathrm{Var}(\psi_j)}{n_t}.$$

Furthermore, by lemma C.1 in the appendix, $\mathrm{Var}(\psi_j) \sim 1$, therefore

$$\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 \sim \frac{|j|}{n_t}.$$

More precisely, proposition C.3 in the appendix and a union bound show that, with probability greater than $1 - e^{-y}$, for any $j \in \{a_x\Delta, \dots, b_x\Delta\}$,

$$\left| \sum_{i=k_*+(j)_-+1}^{k_*+(j)_+} \left(\hat{\theta}_i^T - \theta_i\right)^2 - \frac{|j|}{n_t} \right| \leq \kappa_1(y + \log n + \log((b_x - a_x)\Delta))^2 n^{-\min(\frac{1}{12}, \frac{\delta_4}{2})} \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) \mathfrak{e}.$$

Let $r_n = \kappa_1(y + \log n + \log((b_x - a_x)\Delta))^2 n^{-\min(\frac{1}{12}, \frac{\delta_4}{2})}$. Then for any $j \in \{1, \dots, n_t\}$,

$$\begin{aligned}
\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 &= -\sum_{i=k_*+1}^{k_*+j} \theta_i^2 + \sum_{i=k_*+1}^{k_*+j} \left(\hat{\theta}_i^T - \theta_i\right)^2 \\
&= -\sum_{i=k_*+1}^{k_*+j} \theta_i^2 + \frac{j}{n_t} \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e} \\
&= \sum_{i=k_*+1}^{k_*+j} \left[\frac{1}{n_t} - \theta_i^2\right] \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e} \\
&= \mathfrak{e} f_n\left(\frac{j}{\Delta}\right) \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e}.
\end{aligned}$$

On this same event, for any $j \in \{-k_*(n_t), \dots, -1\}$,

$$\begin{aligned}
\left\|\hat{s}_{k_*+j}^T - s\right\|^2 - \left\|\hat{s}_{k_*}^T - s\right\|^2 &= \sum_{i=k_*+j+1}^{k_*} \theta_i^2 - \sum_{i=k_*+j+1}^{k_*} \left(\hat{\theta}_i^T - \theta_i\right)^2 \\
&= \sum_{i=k_*+j+1}^{k_*} \theta_i^2 - \frac{|j|}{n_t} \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e} \\
&= \sum_{i=k_*+j+1}^{k_*} \left[\theta_j^2 - \frac{1}{n_t}\right] \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e} \\
&= \mathfrak{e} f_n\left(\frac{j}{\Delta}\right) \pm \max\left(1, f_n\left(\frac{j}{\Delta}\right)\right) r_n \mathfrak{e}.
\end{aligned}$$

Thus, since f_n and $\hat{R}_T^{ho}$ are linear between the points of $\frac{1}{\Delta}\mathbb{Z}$,

$$\sup_{\alpha\in\left[\frac{-k_*}{\Delta};\frac{n_t-k_*}{\Delta}\right]}\frac{|L(\alpha)-f_n(\alpha)|}{\max(1,f_n(\alpha))}=\frac{1}{\mathfrak{e}}\max_{-k_*\le j\le n_t-k_*}\frac{\left|\left\|\hat{s}^T_{k_*+j}-s\right\|^2-\left\|\hat{s}^T_{k_*}-s\right\|^2-\mathfrak{e}f_n\left(\frac{j}{\Delta}\right)\right|}{\max(1,f_n\left(\frac{j}{\Delta}\right))}$$
$$\le r_n.$$

This proves claim 5, using lemma C.5 to bound $\log((b_x-a_x)\Delta)$.

B.2 Proof of claim 6

Let $n_v=|T^c|=n-|T|=n-n_t$. Let $F:x\to\int_0^x s(t)dt$ be the cumulative distribution function of the given X_i. Let $F_{T^c}:x\to\frac{1}{n_v}\sum_{i\notin T}\mathbb{I}_{X_i\le x}$ be the empirical cumulative distribution function of the sample $D_n^{T^c}$. By the Komlós-Major-Tusnády approximation theorem [15, Theorem 3], there exist a universal constant C and a standard Brownian bridge process B_{T^c} such that for all $y>0$, with probability greater than $1-e^{-y}$, $\left\|B_{T^c}\circ F-\sqrt{n_v}(F_{T^c}-F)\right\|_\infty\le\frac{C(\log n_v+y)}{\sqrt{n_v}}$ (remark that since F is continuous, $F(X_i)\sim\mathcal{U}([0;1])$, which means that the result for general F follows from the result for the uniform distribution). Furthermore, B_{T^c} can always be realized as a measurable function of $D_n^{T^c}$ and an auxiliary, uniformly distributed random variable ν: $B_{T^c}=H(D_n^{T^c},\nu)$, with ν independant from D_n. Let B^{T^c} be obtained in this way. From $B_{T^c}\circ F$, one can define an operator on the Sobolev space $W^1(\mathbb{R})$:

Definition 10. *For any function f such that $f'\in L^1([0;1])$, let*

$$G_{T^c}(f)=-\int_0^1 f'(x)B_{T^c}(F(x))dx.$$

G_{T^c} "approximates" the empirical process $\sqrt{n_v}(P_n^{T^c}-P)$ on the space W^1. Lemma B.1 below gives a bound on the error made with this approximation.

Lemme B.1. *For any function f such that $f'\in L^1([0;1])$,*

$$\left|G_{T^c}(f)-\sqrt{n_v}(P_n^{T^c}-P)(f)\right|\le\left\|B_{T^c}-\sqrt{n_v}(F_{T^c}-F)\right\|_\infty\left\|f'\right\|_{L^1}.$$

Furthermore, for all functions f,g such that $f',g'\in L^1([0;1])$,

$$\mathrm{Cov}(G_{T^c}(f),G_{T^c}(g))=P[fg]-P[f]P[g]=\mathrm{Cov}\left(\sqrt{n_v}(P_n^{T^c}-P)(f),\sqrt{n_v}(P_n^{T^c}-P)(g)\right).$$

Proof. Let f be a function such that $f'\in L^1([0;1])$. Then

$$\begin{aligned}(P_n^{T^c}-P)(f)&=\int f d(P_n^{T^c}-P)\\&=\int[f-f(0)]d(P_n^{T^c}-P)\\&=\int_0^1\int_0^1\mathbb{I}_{t<x}f'(t)dt\;d(F_{T^c}-F)(x)\\&=\int_0^1 f'(t)(P_n^{T^c}-P)((t,+\infty))\\&=-\int_0^1 f'(t)(F_{T^c}-F)(t)dt.\end{aligned}\tag{34}$$

Il follows that for all functions f such that $f' \in L^1([0;1])$,

$$\begin{aligned}\left|G_{T^c}(f) - \sqrt{n_v}(P_n^{T^c} - P)(f)\right| &= \left|\int_0^1 f'(t)\left[\sqrt{n_v}(F_{T^c} - F) - B_{T^c} \circ F\right](t)dt\right| \\ &\leq \|f'\|_{L^1([0;1])} \left\|B_{T^c} \circ F - \sqrt{n_v}(F_{T^c} - F)\right\|_\infty .\end{aligned}$$

By definition, it is clear that $\mathbb{E}[G_{T^c}(f)] = 0$. Thus,

$$\begin{aligned}\mathrm{Cov}\left(G_{T^c}(f), G_{T^c}(g)\right) &= \mathbb{E}\left[G_{T^c}(f)G_{T^c}(g)\right] \\ &= \mathbb{E}\left[\int_0^1\int_0^1 f'(u)g'(v)B_{T^c}(F(u))B_{T^c}(F(v))\right] \\ &= \int_0^1\int_0^1 f'(u)g'(v)[F(u) \wedge F(v)][1 - F(u) \vee F(v)]dudv \\ &= \int_0^1\int_0^1 f'(u)g'(v)\left(\mathbb{E}[\mathbb{I}_{X\leq u}\mathbb{I}_{X\leq v}] - \mathbb{E}[\mathbb{I}_{X\leq u}]\mathbb{E}[\mathbb{I}_{X\leq v}]\right) \\ &= n_v\int_0^1\int_0^1 f'(u)g'(v)\mathbb{E}\left[(F_{T^c} - F)(u)(F_{T^c} - F)(v)\right] \\ &= \mathrm{Cov}\left(\sqrt{n_v}(P_n^{T^c} - P)(f), \sqrt{n_v}(P_n^{T^c} - P)(g)\right) \text{ by equation (34)}\end{aligned}$$

□

Let the process Z^1 be defined for all integers $j \geq -k_*$ by

$$Z^1\left(\frac{j}{\Delta}\right) = \frac{2}{\sqrt{n_v}\mathfrak{e}}G_{T^c}\left(\hat{s}^T_{k_*+j} - \hat{s}^T_{k_*}\right).$$

Z^1 is extended to the interval $[a;b_x]$ by linear interpolation, as for Z. By lemma B.1, the variance-covariance function of Z^1 conincides with that of Z at the points $\frac{j}{\Delta}, j \in \mathbb{Z} \cap [a;b_x]$, and this property extends by bilinearity to the whole interval $[a;b_x]$. Furthermore,

$$\begin{aligned}\sup_{a_x\leq\alpha\leq b_x}\frac{|Z^1_\alpha - Z_\alpha|}{1 + f_n(\alpha)} &\leq \max_{j\in\mathbb{Z}\cap[a_x\Delta;b_x\Delta]}\frac{\left|Z^1\left(\frac{j}{\Delta}\right) - Z\left(\frac{j}{\Delta}\right)\right|}{1 + f_n\left(\frac{j}{\Delta}\right)} \\ &\leq \frac{4\sqrt{2}\pi}{\mathfrak{e}\sqrt{n_v}}\left\|B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)\right\|_\infty \times \max_{a_x\Delta\leq j\leq b_x\Delta}\frac{\left\|\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} i\hat{\theta}_i^T \sin(2i\pi\cdot)\right\|_1}{1 + f_n\left(\frac{j}{\Delta}\right)} \\ &\leq \frac{4\pi}{\mathfrak{e}\sqrt{n_v}}\left\|B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)\right\|_\infty \times \max_{a_x\Delta\leq j\leq b_x\Delta}\frac{\sqrt{\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+} i^2(\hat{\theta}_i^T)^2}}{1 + f_n\left(\frac{j}{\Delta}\right)}\end{aligned}$$

By construction, the process $B_{T^c} \circ F - \sqrt{n_v}(F_{n_v} - F)$ is independent from D_n^T. As a result,

$$\mathbb{E}\left[\sup_{a\le\alpha\le b_x} \frac{|Z_\alpha^1 - Z_\alpha|}{1+f_n(\alpha)}\Big| D_n^T\right] \le \frac{4\pi}{\mathfrak{e}\sqrt{n_v}}\mathbb{E}\left[\|B_{T^c}\circ F - \sqrt{n_v}(F_{n_v}-F)\|_\infty\right]$$
$$\times (k_* + b_x\Delta) \max_{a_x\Delta\le j\le b_x\Delta} \frac{\sqrt{\sum_{i=k_*+(j)_-+1}^{k_*+(j)_+}\left(\hat\theta_i^T\right)^2}}{1+f_n\left(\frac{j}{\Delta}\right)}$$
$$\le \frac{4\pi C\log n_v}{\mathfrak{e}n_v} \times (k_*+b_x\Delta)\max_{a_x\Delta\le j\le b_x\Delta}\sqrt{\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+}\frac{2\theta_j^2 + 2\left(\hat\theta_j^T-\theta_j\right)^2}{\left(1+f_n\left(\frac{j}{\Delta}\right)\right)^2}}. \quad (35)$$

By proposition C.3 and lemma 5.1, there exists an event $E_1(y)$ of probability greater than $1-e^{-y}$ such that, for all $D_n^T \in E_1(y)$,

$$\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+}\left(\hat\theta_i^T-\theta_i\right)^2 \le \frac{|j|}{n_t} + \kappa_1\frac{|j|}{\Delta}[\log n+y]^2 n^{-\min(\frac{1}{12},\frac{\delta_4}{2})}\mathfrak{e}(n)$$
$$\le \left(1+f_n\left(\frac{j}{\Delta}\right)\right)^2\frac{\Delta}{n_t} + \kappa_1\left(1+f_n\left(\frac{j}{\Delta}\right)\right)^2[\log n+y]^2 n^{-\min(\frac{1}{12},\frac{\delta_4}{2})}\mathfrak{e}(n).$$

Moreover, by lemma A.1,

$$\sum_{i=k_*-(j)_-+1}^{k_*+(j)_+}\theta_i^2 \le 2\left(1+f_n\left(\frac{j}{\Delta}\right)\right)^2\mathcal{E}.$$

It follows that for all $D_n^T \in E_1(y)$,

$$\mathbb{E}\left[\sup_{a_x\le\alpha\le b_x}\frac{|Z_\alpha^1-Z_\alpha|}{1+f_n(\alpha)}\Big|D_n^T\right] \le \frac{4\pi C\log n_v}{\mathfrak{e}n_v}\times(k_*+b_x\Delta)\left[2\sqrt{\mathcal{E}}+\sqrt{2\kappa_1}(\log n+y)n^{-\min(\frac{1}{24},\frac{\delta_4}{4})}\sqrt{\mathfrak{e}}\right].$$

Since $\mathfrak{e}\le\mathcal{E}$ and $n^{-\min(\frac{1}{24},\frac{\delta_4}{4})}\log n\to 0$, there exists therefore a constant κ such that for all $D_n^T\in E_1(y)$:

$$\mathbb{E}\left[\sup_{a_x\le\alpha\le b_x}|Z_\alpha^1-Z_\alpha|\Big|D_n^T\right] \le \kappa\frac{\log n_v}{\sqrt{n_v}}\frac{\sqrt{\mathcal{E}}}{\mathfrak{e}\sqrt{n_v}}\times(k_*+b_x\Delta)(1+y)$$
$$\le \kappa\frac{\log n_v}{\sqrt{n_v}}\times(k_*+b_x\Delta)(1+y). \quad (36)$$

By equation (69) of lemma C.5,

$$k_*+b_x\Delta \le \kappa(1+x)n^{\frac{1}{3}}$$

By hypothesis 6 of section 3,

$$\log n\frac{k_*+b_x\Delta}{\sqrt{n_v}} \le \kappa(1+x)\log n n^{-\frac{\delta_5}{2}}. \quad (37)$$

Since $\frac{\log n}{n^{\frac{\delta_5}{2}}} = o\left(n^{-\frac{\delta_5}{3}}\right)$, by equations (36), (37), there exists a constant $\kappa(c_1,\delta_5)$ such that for any n, with probability greater than $1-e^{-y}$,

$$\mathbb{E}\left[\sup_{a_x\le\alpha\le b_x}\frac{|Z_\alpha^1-Z_\alpha|}{1+f_n(\alpha)}\Big|D_n^T\right] \le \kappa(1+x)(1+y)n^{-\frac{\delta_5}{3}}.$$

B.3 Proof of proposition 6.1

Let $c_{i,j} = \frac{\theta_{i+j}}{\sqrt{2}} + (\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j})\theta_{|i-j|} - \theta_i\theta_j$. U_{I^1,I^2} can be expressed as the sum of 6 terms: $U_{I^1,I^2} = V_1 + V_2 + V_3 + V_4 + V_5 + V_6$, where

$$
\begin{aligned}
V_1 &= \sum_{i\in I^1}\sum_{j\in I^2}\theta_i\theta_j\left[\frac{\theta_{i+j}}{\sqrt{2}} + (\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j})\theta_{|i-j|} - \theta_i\theta_j\right]\\
V_2 &= (P^T - P)\sum_{i\in I^1}\psi_i\sum_{j\in I^2}\theta_j c_{i,j}\\
V_3 &= (P^T - P)\sum_{j\in I^2}\psi_j\sum_{i\in I^1}\theta_i c_{i,j}\\
V_4 &= \frac{1}{\sqrt{2}}\sum_{i\in I^1}\sum_{j\in I^2}(P^T - P)\psi_i(P^T - P)\psi_j\theta_{|i-j|}\\
V_5 &= \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{j\in I^1\cap I^2}\left(\hat{\theta}_j^T - \theta_j\right)^2\\
V_6 &= \sum_{i\in I^1}\sum_{j\in I^2}(P^T - P)\psi_i(P^T - P)\psi_j\left[\frac{\theta_{i+j}}{\sqrt{2}} - \theta_i\theta_j\right]
\end{aligned}
$$

The first term is

$$
V_1 = \left(1 - \frac{1}{\sqrt{2}}\right)\sum_{i\in I^1\cap I^2}\theta_i^2 + \frac{1}{\sqrt{2}}\sum_{i\in I^1}\sum_{j\in I^2}\theta_i\theta_j\theta_{|i-j|} + \sum_{i\in I^1}\sum_{j\in I^2}\theta_i\theta_j\left[\frac{\theta_{i+j}}{\sqrt{2}} - \theta_i\theta_j\right].
$$

For all $i \in I^1$,

$$
\sum_{j\in I^2}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j| \le 2\sum_{j\ge k_* + a_x\Delta + 1}|\theta_j|.
$$

Furthermore, for all $k \ge 2$, by hypothesis 1of section 3

$$
\sum_{j\ge k}|\theta_j| \le \sum_{j=k}^{+\infty}\sqrt{\sum_{i=j}^{+\infty}\theta_i^2} \le \sum_{j=k}^{+\infty}\frac{c_1}{(j-1)^{1+\frac{\delta_1}{2}}} \le \frac{2c_1}{\delta_1}(k-1)^{\frac{-\delta_1}{2}}. \tag{38}
$$

Since $k_* + \alpha_g\Delta \ge \frac{\kappa}{(1+f_n(\alpha_g))^{\frac{1}{\delta_2}}}n_t^{\frac{2}{3\delta_2}}$ by lemma C.5, there is a constant $\kappa(c_1,c_2)$ such that

$$
\sum_{j\in I^2}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j| \le \kappa\frac{(1+x)^{\frac{\delta_1}{2\delta_2}}}{n_t^{\frac{\delta_1}{3\delta_2}}}. \tag{39}
$$

The same argument applies to $\sum_{i\in I^1}\frac{|\theta_{i+j}|}{\sqrt{2}} + |\theta_i||\theta_j|$. Thus, by lemma C.4,

$$
\sum_{i\in I^1}\sum_{j\in I^2}\theta_i\theta_j\left[\theta_{i+j} - \theta_i\theta_j\right] \le 2\kappa\frac{(1+x)^{\frac{\delta_1}{2\delta_2}}}{n_t^{\frac{\delta_1}{3\delta_2}}}\sqrt{\left(\sum_{i\in I^1}\theta_i^2\right)\left(\sum_{j\in I^2}\theta_j^2\right)}.
$$

Thus

$$V_1 = \left(1 - \frac{1}{\sqrt{2}}\right) \sum_{i \in I^1 \cap I^2} \theta_i^2 + \frac{1}{\sqrt{2}} \sum_{i \in I^1} \sum_{j \in I^2} \theta_i \theta_j \theta_{|i-j|} \pm \kappa \frac{(1+x)^{1+\frac{\delta_1}{2\delta_2}}}{n_t^{\frac{\delta_1}{3\delta_2}}} \sqrt{w(I_1) w(I_2)} \mathcal{E} \tag{40}$$

Bernstein's inequality applies to V_2 and V_3. By symmetry, let us only consider V_2. Its variance satisfies the following inequality.

$$\begin{aligned} \operatorname{Var}\left(\sum_{i \in I^1} \psi_i \sum_{j \in I^2} \theta_j c_{i,j}\right) &\leq \|s\|_\infty \left\| \sum_{i \in I^1} \psi_i \sum_{j \in I^2} \theta_j c_{i,j} \right\|^2 \\ &\leq \|s\|_\infty \sum_{i \in I^1} \left(\sum_{j \in I^2} \theta_j c_{i,j} \right)^2 . \end{aligned}$$

Let us now apply lemma C.4. For all $i \in I^1$,

$$\begin{aligned} \sum_{j \in I^2} |c_{i,j}| &\leq \frac{1}{\sqrt{2}} \sum_{j \in I^2} |\theta_{i+j}| + \frac{1}{\sqrt{2}} \sum_{j \in I^2} |\theta_{|i-j|}| + |\theta_i| \sum_{j \in I^2} |\theta_j| \\ &\leq \left(\sqrt{2} + \sup_{i \in \mathbb{N}} |\theta_i|\right) \sum_{r \in \mathbb{N}} |\theta_r| \\ &\leq 3 \|\theta\|_{\ell^1} \end{aligned}$$

In the same way, for all $j \in I^2$, $\sum_{i \in I^1} |c_{i,j}| \leq 3 \|\theta\|_{\ell^1}$, hence by lemma C.4,

$$\begin{aligned} \operatorname{Var}\left(\sum_{i \in I^1} \psi_i \sum_{j \in I^2} \theta_j c_{i,j}\right) &\leq 3 \|\theta\|_{\ell^1} \|s\|_\infty \sum_{j \in I^2} \theta_j^2 \\ &\leq 3 \|\theta\|_{\ell^1} \|s\|_\infty w(I_2) \mathcal{E} \end{aligned} \tag{41}$$

As for the upper bound on the uniform norm, it follows from lemma C.4 and the elementary upper bound $\|\psi_i\|_\infty \leq \sqrt{2}$ that

$$\begin{aligned} \sup_{x \in \mathbb{R}} \left| \sum_{i \in I^1} \psi_i(x) \sum_{j \in I^2} \theta_j c_{i,j} \right| &\leq \sqrt{\sum_{i \in I^1} \left(\sum_{j \in I^2} \theta_j c_{i,j} \right)^2} \sup_{x \in \mathbb{R}} \sqrt{\sum_{i \in I^1} \psi_i(x)^2} \\ &\leq 3 \|\theta\|_{\ell^1} \sqrt{2|I^1|} \sqrt{\sum_{j \in I^2} \theta_j^2} \\ &\leq \kappa \sqrt{w(I_1)\Delta} \sqrt{w(I_2)\mathcal{E}}, \end{aligned} \tag{42}$$

for some constant $\kappa = \kappa(\|\theta\|_{\ell^1})$. By Bernstein's inequality, there exists an event $E_2(y) \subset \mathbb{R}^{n_t}$ with probability $\mathbb{P}(D_n^T \in E_2(y)) \geq 1 - e^{-y}$ such that, for any $D_n^T \in E_2(y)$,

$$\begin{aligned}|V_2| &\leq \sqrt{\frac{2y}{n_t}}\sqrt{\operatorname{Var}\left(\sum_{i\in I^1}\psi_i\sum_{j\in I^2}\theta_j c_{i,j}\right)} + \frac{y}{3n_t}\sup_{x\in\mathbb{R}}\left|\sum_{i\in I^1}\psi_i(x)\sum_{j\in I^2}\theta_j c_{i,j}\right| \\ &\leq \sqrt{6\|\theta\|_{\ell^1}\|s\|_\infty w(I_2)}\sqrt{\frac{y\mathcal{E}}{n_t}} + \frac{\kappa y}{3n_t}\sqrt{w(I_1)w(I_2)}\sqrt{\Delta\mathcal{E}} \text{ by (41), (42).}\end{aligned}$$

Setting $\kappa = \max\left(\sqrt{6\|\theta\|_{\ell^1}\|s\|_\infty}, \frac{\kappa}{3}\right)$, it follows that on $E_2(y)$,

$$|V_2| \leq \kappa\sqrt{yw(I_2)}\sqrt{\frac{n_v}{n_t}}\mathfrak{e} + \kappa y\sqrt{w(I_1)w(I_2)}\sqrt{\frac{n_v}{n_t}}\mathfrak{e}\sqrt{\mathcal{E}}.$$

By lemma 5.3, $\mathfrak{e} \leq \mathcal{E}$ and $\mathcal{E}$ is uniformly bounded: $\mathcal{E} \leq 2\sum_{j=1}^{n_t}\theta_j^2 + \frac{1}{n-n_t} \leq 1 + 2\|s\|^2 \leq 1 + 2\|s\|_\infty$. Furthermore, by hypothesis 5of section 3, $\sqrt{\frac{n_v}{n_t}} = \sqrt{\frac{n-n_t}{n_t}} \leq n^{-\frac{\delta_4}{2}}$. Since $w(I_1) \leq 1$, there exists a constant $\kappa(\|\theta\|_{\ell^1}, \|s\|_\infty)$ such that, on $E_2(y)$,

$$|V_2| \leq \kappa y\sqrt{w(I_1)w(I_2)}n^{-\frac{\delta_4}{2}}\mathfrak{e}. \tag{43}$$

Symmetrically, there exists an event $E_3(y)$ of probability greater than $1-e^{-y}$, such that for any $D_n^T \in E_3(y)$,

$$|V_3| \leq \kappa y\sqrt{w(I_1)w(I_2)}n^{-\frac{\delta_4}{2}}\mathfrak{e}. \tag{44}$$

Now consider V_4. This term can be expressed as a finite sum of sums of squares:

$$\begin{aligned}V_4 &= \frac{1}{\sqrt{2}}\sum_{r\in\mathbb{Z}}\sum_{i\in I^1\cap(I^2-r)}(P^T-P)\psi_i(P^T-P)\psi_{i+r}\theta_{|r|} \\ &= \frac{1}{4\sqrt{2}}\sum_{r\in\mathbb{Z}}\theta_{|r|}\sum_{i\in I^1\cap(I^2-r)}\left[(P^T-P)(\psi_i+\psi_{i+r})\right]^2 - \left[(P^T-P)(\psi_i-\psi_{i+r})\right]^2.\end{aligned}$$

Let $J_0 = \{j \in \mathbb{N} : \lfloor\frac{j}{r}\rfloor \text{ is even}\}$ and $J_1 = \{j \in \mathbb{N} : \lfloor\frac{j}{r}\rfloor \text{ is odd}\}$. Thus

$$V_4 = \frac{1}{4\sqrt{2}}\sum_{r\in\mathbb{Z}}\theta_{|r|}\sum_{(z,\varepsilon)\in\{0;1\}\times\{-1;1\}}\sum_{j\in J_z}\varepsilon(P^T-P)(\psi_i+\varepsilon\psi_{i+r})^2\mathbb{I}_{I^1}(i)\mathbb{I}_{I^2}(i+r).$$

For any fixed $r \neq 0$, $(z,\varepsilon) \in \{0;1\}\times\{-1;1\}$, $\frac{1}{\sqrt{2}}\left(\psi_i+\varepsilon\psi_{i+r}\right)_{i\in J_z}$ is an orthonormal collection of functions, since for any $(i,j) \in J_z^2$,

$$\begin{aligned}<\psi_i+\varepsilon\psi_{i+r}, \psi_j+\varepsilon\psi_{j+r}> &= <\psi_i,\psi_j> + \varepsilon<\psi_{i+r},\psi_j> + \varepsilon<\psi_i,\psi_{j+r}> + <\psi_{i+r},\psi_{j+r}> \\ &= 2\delta_{i,j} + \varepsilon<\psi_{i+r},\psi_j> + \varepsilon<\psi_i,\psi_{j+r}> \\ &= 2\delta_{i,j} \text{ since } i,j \in J_z \text{ and } i+r, j+r \in J_{1-z}.\end{aligned}$$

[2, Lemma 14] applied to $S_m = \langle (\psi_i + \varepsilon\psi_{i+r})_{i\in J_z\cap I^1\cap I^2}\rangle$ for all $(z,\varepsilon)\in\{0;1\}\times\{-1;1\}$, $r\in\{-n_t,\dots,n_t\}$ and a union bound yield an event $E_4(y)$ of probability $\mathbb{P}(D_n^T\in E_4(y))\geq 1-e^{-y}$ such that, for some absolute constant κ and for all $D_n^T\in E_4(y)$, $(z,\varepsilon)\in\{0;1\}\times\{-1;1\}$ and $r\in\mathbb{Z}$,

$$\begin{aligned}\sum_{i\in J_z\cap I^1\cap(I^2-r)}\varepsilon(P^T-P)(\psi_i+\varepsilon\psi_{i+r})^2 = (1\pm\delta)\frac{\varepsilon}{n_t}\sum_{i\in J_z\cap I^1\cap(I^2-r)}\big[\mathrm{Var}(\psi_i)+\mathrm{Var}(\psi_{i+r})\\ +2\varepsilon\,\mathrm{Cov}(\psi_i,\psi_{i+r})\big]+\kappa\frac{\|s\|_\infty[\log(1+r)+\log n_t+y]}{(\delta\wedge 1)n_t}\\ +\kappa\frac{|I^1\cap I^2|[\log(1+r)+\log n_t+y]^2}{(\delta\wedge 1)^3 n_t^2}.\end{aligned}$$

By summing on $(r,z,\varepsilon)\in\mathbb{Z}\times\{0;1\}\times\{-1;1\}$ and since $\|\psi_i\|_\infty\leq\sqrt{2}$, it follows that for all $D_n^T\in E_4(y)$,

$$\begin{aligned}\left|V_4-\frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I^1\cap(I^2-r)}c_{i,i+r}\right| &= \left|V_4-\frac{1}{n_t}\sum_{r=-n_t}^{n_t}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I^1\cap(I^2-r)}c_{i,i+r}\right|\\ &\leq \frac{\delta}{n_t\sqrt{2}}\sum_{r\in\mathbb{Z}}|\theta_{|r|}|\sum_{i\in I^1\cap(I^2-r)}[\mathrm{Var}(\psi_i)+\mathrm{Var}(\psi_{i+r})]\\ &\quad+\kappa\sum_{r\in\mathbb{Z}}\frac{|\theta_{|r|}|}{\sqrt{2}}\times\frac{\|s\|_\infty[\log n_t\log(1+r)+y]}{(\delta\wedge 1)n_t}\\ &\quad+\kappa\sum_{r\in\mathbb{Z}}\frac{|\theta_{|r|}|}{\sqrt{2}}\times\frac{|I^1\cap I^2|[\log n_t+\log(1+r)+y]^2}{(\delta\wedge 1)^3 n_t^2}\end{aligned}$$

It follows from hypothesis 1that the sum $\sum_{r\in\mathbb{Z}}|\theta_{|r|}|\log(1+r)^2$ converges to a finite value $\|\theta\|_{1,\log^2}$. Thus,

$$\begin{aligned}\left|V_4-\frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I^1\cap(I^2-r)}c_{i,i+r}\right| &\leq 3\|\theta\|_{\ell^1}\min(w(I^1),w(I^2))\frac{\delta}{n_t}\Delta+2\kappa\|\theta\|_{1,\log^2}\frac{\|s\|_\infty[1+y]}{(\delta\wedge 1)n_t}\\ &\quad+8\kappa\min(w(I^1),w(I^2))\|\theta\|_{1,\log^2}\frac{\Delta[1+y]^2}{(\delta\wedge 1)^3 n_t^2}.\end{aligned}$$

There exists therefore a constant $\kappa(\|\theta\|_{1,\log^2})$ such that, for all $D_n^T\in E_4(y)$,

$$\left|V_4-\frac{1}{n_t}\sum_{r\in\mathbb{Z}}\frac{\theta_{|r|}}{\sqrt{2}}\sum_{i\in I^1\cap(I^2-r)}c_{i,i+r}\right|\leq\kappa\delta\min(w(I^1),w(I^2))\mathcal{E}+\frac{[\log n_t+y]}{(\delta\wedge 1)n_t}+\kappa\min(w(I^1),w(I^2))\frac{[\log n_t+y]^2}{(\delta\wedge 1)^3 n_t}\mathcal{E}.$$

Let now $\delta=\max\left\{\frac{n-n_t}{n_t},n^{-\frac{1}{3}}\right\}^{\frac{3}{4}}$. By hypothesis 5 of section $3\frac{n-n_t}{n_t}\leq n^{-\delta_4}$, therefore $\delta\mathcal{E}\leq n^{-\min\left(\frac{1}{4},\frac{3\delta_4}{4}\right)}\mathcal{E}$. Moreover, $\mathcal{E}\geq\frac{1}{n_v}$ therefore $\frac{1}{\delta n_t}\leq\left(\frac{n-n_t}{n_t}\right)^{\frac{1}{4}}\frac{1}{n_v}\leq n^{-\frac{\delta_4}{4}}\mathcal{E}$. Finally, since $\delta\geq n^{-\frac{1}{4}}$ and $n_t\geq\frac{n}{2}$, $\frac{\mathcal{E}}{\delta^3 n_t}\leq 2n^{-\frac{1}{4}}\mathcal{E}$. Since $\delta_4\leq 1$, there exists therefore a constant κ such that for all $D_n^T\in E_4(y)$,

$$\left|V_4-\frac{1}{n_t\sqrt{2}}\sum_{i\in I^1}\sum_{j\in I^2}\theta_{|i-j|}c_{i,j}\right|\leq\kappa\min(w(I^1),w(I^2))[\log n_t+y]^2 n^{-\frac{\delta_4}{4}}\mathcal{E}. \tag{45}$$

Moreover, since $c_{i,j} = \frac{\theta_{i+j}}{\sqrt{2}} + \left(\frac{1-\delta_{i,j}}{\sqrt{2}} + \delta_{i,j}\right)\theta_{|i-j|} - \theta_i\theta_j$ and $\theta_0 = 1$,

$$\frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}}{\sqrt{2}}c_{i,j} = \sum_{i\in I^1\cap I^2}\frac{1}{\sqrt{2}}\left(1-\frac{1}{\sqrt{2}}\right)\frac{1}{n_t} + \frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}^2}{2} + \frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}\theta_{i+j}}{2} - \frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}}{\sqrt{2}}\theta_i\theta_j.$$

Since for all $j \in \mathbb{N}$, $|\theta_j| \leq 1$,

$$\begin{aligned}\left|\frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}}{\sqrt{2}}c_{i,j} - \left(1-\frac{1}{\sqrt{2}}\right)\frac{|I^1\cap I^2|}{n_t\sqrt{2}} - \frac{1}{n_t}\sum_{i\in I^1}\sum_{j\in I^2}\frac{\theta_{|i-j|}^2}{2}\right| &\leq \frac{2}{n_t}\left(\sum_{r\in\mathbb{N}}|\theta_r|\right)^2 \\ &\leq 2\frac{n-n_t}{n_t}\frac{1}{n_v}\|\theta\|_{\ell^1}^2 \\ &\leq 2\|\theta\|_{\ell^1}^2 n^{-\delta_4}\mathcal{E} \qquad (46)\end{aligned}$$

since $\mathcal{E} \geq \frac{1}{n_v}$ and $\frac{n-n_t}{n_t} \geq n^{-\delta_4}$, by hypothesis 5of section 3From equations (45) and (46), it follows that, for some constant $\kappa(\|\theta\|_{1,\log^2})$,

$$\left|V_4 - \left(1-\tfrac{1}{\sqrt{2}}\right)\frac{|I^1\cap I^2|}{n_t\sqrt{2}} - \frac{1}{2n_t}\sum_{i\in I^1}\sum_{j\in I^2}\theta_{|i-j|}^2\right| \leq \kappa\min(w(I^1),w(I^2))[\log n_t + y]^2 n^{-\frac{\delta_4}{4}}\mathcal{E}. \qquad (47)$$

V_5 can be expressed as

$$V_5 = \left(1-\frac{1}{\sqrt{2}}\right)\sum_{j\in I^1\cap I^2}\left(\hat{\theta}_j^T - \theta_j\right)^2,$$

therefore by proposition C.3, there exists an event $E_5(y)$ of probability greater than $1-e^{-y}$ such that for all $D_n^T \in E_5(y)$,

$$\left|V_5 - \left(1-\frac{1}{\sqrt{2}}\right)\frac{|I^1\cap I^2|}{n_t}\right| \leq \kappa_1\min(w(I^1),w(I^2))[\log n + y]^2 n^{-\min(\frac{1}{4},\frac{\delta_4}{2})}\mathcal{E}. \qquad (48)$$

Finally, by lemma C.4 and equation (39),

$$V_6 \leq \kappa\frac{(1+x)^{\frac{\delta_1}{2\delta_2}}}{n_t^{\frac{\delta_1}{3\delta_2}}}\sqrt{\left(\sum_{i\in I^1}(P^T-P)^2\psi_i\right)\left(\sum_{j\in I^2}(P^T-P)^2\psi_j\right)} \qquad (49)$$

By proposition C.3, there exists an event $E_6(y)$ of probability greater than $1-e^{-y}$, such that for any $D_n^T \in E_6(y)$,

$$\sum_{j\in I^1}\left(\hat{\theta}_j^T-\theta_j\right)^2 \leq \frac{|I^1|}{\Delta}\mathcal{E} + \kappa_1 w(I^1)(y+\log n)^2 n^{-\min(\frac{1}{4},\frac{\delta_4}{2})}\mathcal{E}$$

$$\sum_{j\in I^2}\left(\hat{\theta}_j^T-\theta_j\right)^2 \leq \frac{|I^2|}{\Delta}\mathcal{E} + \kappa_1 w(I^2)(y+\log n)^2 n^{-\min(\frac{1}{4},\frac{\delta_4}{2})}\mathcal{E}.$$

It follows by equation (49) that on $E_6(y)$, for a certain constant $\kappa(\kappa_1, \delta_1, c_1, \kappa_6)$,

$$V_6 \leq \kappa \sqrt{w(I^1)w(I^2)}[y + \log n]^2 \frac{(1+x)^{\frac{\delta_1}{2\delta_2}}}{n_t^{\frac{\delta_1}{3\delta_2}}} \mathcal{E}. \tag{50}$$

Combining equations (40), (43), (44), (47), (48), (50) on the event $\cap_{i=2}^6 E_i(\log 6 + y)$ yields the result.

B.4 Proof of claim 7

Assume without loss of generality that $m_1 < m_2 < m_3$. We start by proving equation (23). First, changing variables from j_1, j_2 to $i = j_1, r = j_2 - j_1$ yields

$$\begin{aligned}
\sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} &= \sum_{r\in\mathbb{N}} \theta_r \sum_{i=m_2+1-r}^{m_2} \mathbb{I}_{i\geq m_1+1}\mathbb{I}_{i+r\leq m_3}\theta_i\theta_{i+r} \\
&\leq \sum_{r\leq r_0} |\theta_r| \sum_{i=(m_2+1-r)\vee(m_1+1)}^{m_2\wedge(m_3-r)} |\theta_i||\theta_{i+r}| \\
&\quad + \sum_{r>r_0} |\theta_r| \sqrt{\left(\sum_{j_1=m_1+1}^{m_2} \theta_{j_1}^2\right)\left(\sum_{j_2=m_2+1}^{m_3} \theta_{j_2}^2\right)} \\
&\leq \|\theta\|_{\ell^1} \max_{1\leq r\leq r_0} \sum_{i=(m_2+1-r)\vee(m_1+1)}^{m_2\wedge(m_3-r)} |\theta_i||\theta_{i+r}| \\
&\quad + \mathcal{E} \sum_{r>r_0} |\theta_r| \sqrt{w_0(k_1-k_*)w_0(k_2-k_*)}.
\end{aligned} \tag{51}$$

Let us now bound $|\theta_i\theta_{i+r}|$ where $i \in [m_1+1, \ldots, m_2]$ and $i+r \in [m_2+1, \ldots, m_3]$. Necessarily, $k_1 \in \{m_1, m_2\}$ and $k_2 \in \{m_2, m_3\}$. Moreover, $|k_1 - k_*| \geq |i - k_*|$ and $|k_2 - k_*| \geq |i + r - k_*|$. It follows from claim 9 that

$$\begin{aligned}
\theta_i^2 &\leq c_3c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t} + \kappa(1+x)^{\frac{\delta_6}{\delta_2}} w_0(k_1-k_*)n^{-u_2}\mathcal{E} \\
\theta_{i+r}^2 &\leq c_3c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t} + \kappa(1+x)^{\frac{\delta_6}{\delta_2}} w_0(k_2-k_*)n^{-u_2}\mathcal{E}
\end{aligned}$$

which implies that

$$|\theta_i||\theta_{i+r}| \leq \kappa \sqrt{w_0(k_1-k_*)w_0(k_2-k_*)} \max\left(\frac{k_*^{\delta_3-\delta_6}}{n_t}, (1+x)^{\frac{\delta_6}{\delta_2}} n^{-u_2}\mathcal{E}\right)$$

for some constant κ. This yields

$$\max_{1\leq r\leq r_0} \sum_{i=(m_2+1-r)\vee(m_1+1)}^{m_2\wedge(m_3-r)} |\theta_i\theta_{i+r}| \leq \kappa r_0 \sqrt{w_0(k_1-k_*)w_0(k_2-k_*)} \max\left(\frac{k_*^{\delta_3-\delta_6}}{n_t}, (1+x)^{\frac{\delta_6}{\delta_2}} n^{-u_2}\mathcal{E}\right). \tag{52}$$

for any integer r_0. On the other hand, by hypothesis 1 of section 3,

$$\begin{aligned}
\sum_{j=r_0+1}^{+\infty} |\theta_j| &= \sum_{j=0}^{+\infty} \sum_{i=2^j r_0+1}^{2^{j+1} r_0} |\theta_j| \\
&\leq \sum_{j=0}^{+\infty} 2^{j/2} \sqrt{\sum_{i=2^j r_0+1}^{+\infty} \theta_i^2} \\
&\leq \sum_{j=0}^{+\infty} 2^{j/2} \sqrt{r_0} \sqrt{c_1 (2^j r_0)^{-(2+\delta_1)}} \\
&\leq \sqrt{\frac{c_1}{r_0}} \sum_{j=0}^{+\infty} 2^{-\frac{j}{2}} \\
&\leq (2+\sqrt{2}) \sqrt{\frac{c_1}{r_0}}.
\end{aligned}$$

Combining equations (52) and (51) and optimizing over r_0 yields

$$\sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{j_1} \theta_{j_2} \theta_{|j_1-j_2|} \leq \kappa \sqrt{w_0(k_1-k_*) w_0(k_2-k_*)} \left(\mathcal{E}^{\frac{2}{3}} \left(\frac{k_*^{\delta_3-\delta_6}}{n_t} \right)^{\frac{1}{3}} + (1+x)^{\frac{\delta_6}{3\delta_2}} n^{-\frac{2u_2}{3}} \mathcal{E} \right) \tag{53}$$

for some constant $\kappa > 0$. This proves equation (23).

Moreover,

$$\begin{aligned}
\sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{|j_1-j_2|}^2 &= \sum_{r \in \mathbb{N}} \theta_r^2 \left| \{ j_1 : (m_1+1 \leq j_1 \leq m_2) \wedge (m_2+1 \leq j_1+r \leq m_3) \} \right| \\
&\leq \sum_{r \in \mathbb{N}} \theta_r^2 \min(r, m_2-m_1, m_3-m_2) \\
&\leq r_0 \sum_{r=0}^{r_0} \theta_r^2 + \min(m_2-m_1, m_3-m_2) \sum_{r>r_0} \theta_r^2 \\
&\leq r_0 \|s\|^2 + \min(|k_2-k_*|, |k_1-k_*|) \sum_{r>r_0} \theta_r^2.
\end{aligned}$$

by hypothesis 1 of Theorem 1. Let now $r_0 = \lceil \Delta^{\frac{1}{3}} \rceil$. Since $\Delta \geq 1$, it follows that:

$$\begin{aligned}
\frac{1}{n_t} \sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{|j_1-j_2|}^2 &\leq \frac{\Delta^{\frac{1}{3}}+1}{n_t} \|s\|^2 + 2c_1 \min(w_0(k_1-k_*), w_0(k_2-k_*)) \frac{\Delta}{n_t} (\Delta)^{\frac{-2}{3}} \\
&\leq \left[2 \frac{\|s\|^2}{(\Delta)^{\frac{2}{3}}} \mathcal{E} + 2c_1 \min(w_0(k_1-k_*), w_0(k_2-k_*)) \mathcal{E} (\Delta)^{-\frac{2}{3}} \right] \\
&\leq \left[2 \|s\|^2 + 2c_1 \min(w_0(k_1-k_*), w_0(k_2-k_*)) \right] \frac{\mathcal{E}}{(\Delta)^{\frac{2}{3}}}.
\end{aligned}$$

On the other hand, $\Delta \geq \frac{n_t}{n-n_t} \geq n^{\delta_4}$ by hypothesis 5 of Theorem 1. There exists therefore $\kappa(c_1, \|s\|^2)$ such that, for any n,

$$\frac{1}{n_t} \sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta^2_{|j_1-j_2|} \leq \kappa \min(w_0(k_1-k_*), w_0(k_2-k_*)) n^{-\frac{2\delta_4}{3}} \mathcal{E}, \tag{54}$$

which proves equation (24). By proposition 6.1 and the fact that

$$\begin{aligned} w([m_1+1..m_2]) &\leq w_0(k_1-k_*) \\ w([m_2+1..m_2]) &\leq w_0(k_2-k_*), \end{aligned}$$

there exists an event A of probability greater than $1-e^{-y}$ and a constant κ such that

$$\begin{aligned} |E_{m_1,m_2,m_3}| \leq \frac{1}{\sqrt{2}} \sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta_{j_1}\theta_{j_2}\theta_{|j_1-j_2|} + \frac{1}{2n_t} \sum_{j_1=m_1+1}^{m_2} \sum_{j_2=m_2+1}^{m_3} \theta^2_{|j_1-j_2|} \\ + \kappa(y+\log n)^2(1+x)^{\frac{3}{2}} n^{-u_3} \sqrt{w_0(k_1-k_*)w_0(k_2-k_*)}\mathcal{E}. \end{aligned} \tag{55}$$

From equations (55), (53) and (54), equation (25) follows with $u_4 = \min\left(u_3, \frac{2\delta_4}{3}, \frac{2u_2}{3}\right)$.

B.5 End of the proof

Let

$$r = \frac{k_*^{\delta_3-\delta_6}}{\Delta}, \varepsilon = \kappa(1+x)^{u_6}[y+\log n]^2 \left(r + r^{\frac{1}{3}} + n^{-u_5 \wedge u_2 \wedge \delta_4}\right)$$

for some large enough κ. Thus, by Proposition 6.2 above applied on the event E_y conditionally on D_n^T, there exists a continuous Gaussian process $Z^2(D_n^T)$ with variance-covariance function $K(g_n)$ and such that for some constant $\kappa \geq 0$ and all $D_n^T \in E_y$,

$$\mathbb{E}\left[\sup_{a_x \leq \alpha \leq b_x} \frac{|Z^1(\alpha) - Z^2(\alpha)|}{\sqrt{1+|h(\alpha)|}} \Big| D_n^T\right] \leq \sqrt{\log\left(1 + (b_x-a_x)\Delta + |h(a_x)| \vee |h(b_x)|\right)} \varepsilon^{\frac{1}{12}}.$$

Using the facts that $\mathcal{E} \geq \frac{1}{n-n_t}$,

$$h(\alpha) = 4\|s\|^2 \alpha + 8\|s\|_\infty h_0(\alpha) = (4\|s\|^2 + 8\|s\|_\infty)\alpha - 8\|s\|_\infty \frac{\mathfrak{e}}{\mathcal{E}} f_n(\alpha)$$

and the definition of a_x, b_x yields

$$\begin{aligned} |h(a_x)| \vee |h(b_x)| &\leq 12\|s\|_\infty (|a_x| \vee |b_x|) + \frac{8\|s\|_\infty}{\mathcal{E}} \max(\mathfrak{e} f_n(a_x), \mathfrak{e} f_n(n_x)) \\ &\leq 12\|s\|_\infty (|a_x| \vee |b_x|) + 8x\|s\|_\infty (2(n-n_t)or(n_t)+1) \\ &\leq 12\|s\|_\infty (b_x - a_x) + 8x\|s\|_\infty (2n_t or(n_t)+1). \end{aligned}$$

By lemma C.5

$$(b_x-a_x)\Delta + |h(b_x)| \vee |h(a_x)| \leq \kappa(1+x)n^{\frac{1}{3}}$$

for some constant κ. To conclude, remark that by equation (28),

$$\sup_{\alpha\in[a_x;b_x]}\frac{\sqrt{1+|h(\alpha)|}}{1+f_n(\alpha)}\leq\sqrt{3(4\left\|s\right\|^2\vee 8\left\|s\right\|_\infty)}.$$

Thus, for some constants κ and any $u\leq u_5\wedge u_2\wedge\delta_4$,

$$\forall y>0,\forall D_n^T\in E_y,\mathbb{E}\left[\sup_{a_x\leq t\leq b_x}\frac{|Z^1(t)-Z^2(t)|}{1+f_n(t)}|D_n^T\right]\leq\kappa(1+x)^{\frac{u_6}{12}}[y+\log n]^{\frac{1}{6}}\left(r+r^{\frac{1}{3}}+n^{-u}\right)^{\frac{1}{12}}. \tag{56}$$

Since the conditional distribution of $Z^2(D_n^T)$ given D_n^T is entirely determined by the function g_n which does not depend on D_n^T, Z^2 is independent from D_n^T. Moreover, since g_n increases, $W=Z^2\circ g_n^{-1}$ is a continuous, centered gaussian process with covariance function

$$\mathrm{Cov}(Z_s,Z_t)=K(g_n)(g_n^{-1}(s),g_n^{-1}(t))=\begin{cases}s\wedge t\text{ if }0\leq s,t\\-(s\vee t)\text{ if }s,t\leq 0\\0\text{ else }\end{cases}, \tag{57}$$

it is therefore a two-sided Wiener process on $[g_n(a_x);g_n(b_x)]$ taking value 0 at 0. W can be extended to $\mathbb{R}$ by placing independent Wiener processes W_g,W_d on its left and on its right, by the equations $W(u)=W(g_n(a_x))+W_g(u)-W_g(g_n(a_x))$ for $u<a_x$, $W(u)=W(g_n(b_x))+W_d(u)-W_d(g_n(b_x))$ for $u>b_x$. Thus, by claim 6 and equation (56), with probability greater than $1-2e^{-y}$,

$$\begin{aligned}\mathbb{E}\left[\sup_{a_x\leq t\leq b_x}\frac{|Z(t)-W_{g_n(t)}|}{1+f_n(t)}|D_n^T\right]&=\mathbb{E}\left[\sup_{a_x\leq t\leq b_x}\frac{|Z(t)-Z^2(t)|}{1+f_n(t)}|D_n^T\right]\\&\leq\mathbb{E}\left[\sup_{a_x\leq t\leq b_x}\frac{|Z^1(t)-Z(t)|}{1+f_n(t)}|D_n^T\right]+\mathbb{E}\left[\sup_{a_x\leq t\leq b_x}\frac{|Z^1(t)-W_{g_n(t)}|}{1+f_n(t)}|D_n^T\right]\\&\leq\kappa_5(1+x)(1+y)n^{-\frac{\delta_5}{3}}+\kappa(1+x)^{\frac{u_6}{12}}[y+\log n]^{\frac{1}{6}}\left(r+r^{\frac{1}{3}}+n^{-u_5\wedge u_2\wedge\delta_4}\right)^{\frac{1}{12}}\\&\leq\kappa(1+y)(1+x)^{u'}\left(n^{-u}+(\log n)^{\frac{1}{6}}r^{\frac{1}{36}}(r\vee 1)^{\frac{1}{18}}\right),\end{aligned}$$

for all $u<\min\left(\frac{u_5}{12},\frac{u_2}{12},\frac{\delta_4}{12},\frac{\delta_5}{3}\right)$, $u'\geq\max\left(1,\frac{u_6}{12}\right)$ and a constant $\kappa(u)$. Finally, by claim 5, with probability greater than $1-3e^{-y}$,

$$\begin{aligned}&\mathbb{E}\left[\sup_{\alpha\in[a_x;b_x]}\frac{\left|\hat{R}_T^{ho}(\alpha)-[f_n(\alpha)-W_{g_n(\alpha)}]\right|}{1+f_n(\alpha)}\right]\\&\leq\mathbb{E}\left[\sup_{\alpha\in[a_x;b_x]}\frac{|L(\alpha)-f_n(\alpha)|}{1+f_n(\alpha)}\right]+\mathbb{E}\left[\sup_{\alpha\in[a_x;b_x]}\frac{|Z(\alpha)-W_{g_n(\alpha)}|}{1+f_n(\alpha)}\right]\\&\leq\kappa_1(1+x)[\log(n)+\log(2+x)+y]^2n^{-\min(\frac{1}{12},\frac{\delta_4}{2})}+\kappa(1+y)(1+x)^{u'}\left(n^{-u}+(\log n)^{\frac{1}{6}}r^{\frac{1}{36}}(r\vee 1)^{\frac{1}{18}}\right)\\&\leq\kappa(1+y)^2(1+x)^{u_7}\left(n^{-u_1}+(\log n)^{\frac{1}{6}}r^{\frac{1}{36}}(r\vee 1)^{\frac{1}{18}}\right),\end{aligned}$$

for any $u_1<\min\left(\frac{u_5}{12},\frac{u_2}{12},\frac{\delta_4}{12},\frac{\delta_5}{3}\right)$, any $u_7\geq\max\left(1,\frac{u_6}{12}\right)$ and a constant κ. This proves Theorem 1.

C Auxiliary results

Lemme C.1. *Let X be a random variable belonging to $[-1;1]$, with pdf s. For all $j \in \mathbb{N}$, let $\theta_j = \langle s, \psi_j \rangle$. Then*

$$\mathrm{Var}\left(\psi_j(X)\right) \underset{j\to+\infty}{\longrightarrow} 1$$

$$\forall k_0 \leq k, \sum_{j=k_0}^{k} |\mathrm{Var}(\psi_j) - 1| \leq \|\theta\|_{\ell^1} = \sum_{j=0}^{+\infty} |\langle s, \psi_j \rangle|.$$

Proof. $\mathbb{E}[\psi_j(X)] = \int_0^1 \psi_j(x)s(x)dx = \theta_j$. Moreover, $\psi_j(X)^2 = 2\cos^2(2\pi jX) = 1 + \cos(2\pi jX)$, therefore

$$\mathrm{Var}(\cos(\pi jX)) = 1 + \frac{\theta_j}{\sqrt{2}} - \theta_j^2,$$

therefore since $|\theta_j| \leq \sqrt{2}$, $|\mathrm{Var}(\cos(jX)) - 1| \leq \left|\sqrt{2} - \frac{1}{\sqrt{2}}\right| |\theta_j| \leq |\theta_j|$. □

Lemme C.2. *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a function, $g, h : \mathbb{R}_+ \to \mathbb{R}$ be two non-increasing functions. Then*

$$\inf_{x\in\mathbb{R}_+} \{f(x) + g(x) + h(x)\} \leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \inf_{x\in\mathbb{R}_+} \{f(x) + h(x)\}.$$

Proof. Let $\delta > 0$. Let x_g be such that $f(x_g) + g(x_g) \leq \delta + \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\}$. Let x_h be such that $f(x_h) + h(x_h) \leq \inf_{x\in\mathbb{R}_+} \{f(x) + h(x)\}$. Let $x_* = \max(x_g, x_h)$. If $x_* = x_g$, then

$$\begin{aligned}
f(x_*) + g(x_*) + h(x_*) &\leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \delta + h(x_*) \\
&\leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \delta + h(x_h) \text{ since } h \text{ is non-increasing} \\
&\leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \delta + f(x_h) + h(x_h) \\
&\leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \inf_{x\in\mathbb{R}_+} \{f(x) + h(x)\} + 2\delta
\end{aligned}$$

Symetrically, if $x_* = x_h$, then $f(x_*) + g(x_*) + h(x_*) \leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} + \inf_{x\in\mathbb{R}_+} \{f(x) + h(x)\} + 2\delta$. As a result,

$$\begin{aligned}
\inf_{x\in\mathbb{R}_+} \{f(x) + g(x) + h(x)\} \leq f(x_*) + g(x_*) + h(x_*) \leq \inf_{x\in\mathbb{R}_+} \{f(x) + g(x)\} \\
+ \inf_{x\in\mathbb{R}_+} \{f(x) + h(x)\} + 2\delta.
\end{aligned}$$

Since no assumptions were made about $\delta > 0$, lemma C.2 is proved. □

Proposition C.3. *For any integers $k_0 \leq k$, with probability greater than $1 - e^{-y}$:*

$$\begin{aligned}
\left| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t} \right| &\leq \frac{3\|\theta\|_{\ell^1}}{n_t} + (1+\kappa)\|s\|_\infty (y + \log n) \\
&\times \left[\frac{\sqrt{|k-k_0|}}{n_t} + (y + \log n) \frac{|k-k_0|}{n_t^{\frac{5}{4}}} \right]. \qquad (58)
\end{aligned}$$

In particular, there exists a constant $\kappa_1 = \kappa_1(\|s\|_\infty, c_1, \|\theta\|_{\ell^1})$ *such that for any* α_1, α_2 *such that* $(\alpha_1\Delta, \alpha_2\Delta) \in \mathbb{N}^2$ *and* $\alpha_1 < \alpha_2$*, with probability greater than* $1 - e^{-y}$*,*

$$\left| \sum_{j=k_*+\alpha_1\Delta}^{k_*+\alpha_2\Delta} (\hat{\theta}_j^T - \theta_j)^2 - [\alpha_2 - \alpha_1]\mathcal{E} \right| \leq \kappa_1 \max(1, |\alpha_2 - \alpha_1|) [\log n + y]^2 n^{-\min(\frac{1}{4}, \frac{\delta_4}{2})} \mathcal{E} \tag{59}$$

$$\leq \kappa_1 [\max(1, f_n(\alpha_1)) + \max(1, f_n(\alpha_2))] [\log n + y]^2 \times n^{-\min(\frac{1}{12}, \frac{\delta_4}{2})} \mathfrak{e}(n). \tag{60}$$

Proof. Let $(k_0, k) \in \mathbb{N}^2$ be such that $k_0 < k$. The proof rests on lemma 14 of [2] applied to $S_m = \langle \psi_{k_0+1}, \dots, \psi_k \rangle$. Let us compute $b_m = \sup_{u \in \mathbb{R}^{|k-k_0|}: \|u\| \leq 1} \sum_{j=k_0}^k u_j \psi_j(x) \leq \sup_x \sqrt{\sum_{j=k_0}^k \psi_j^2(x)} \leq \sqrt{|k - k_0|}$ and

$$\mathcal{D}_k = \sum_{j=k_0+1}^{k} \mathrm{Var}(\psi_j(X)) = |k - k_0| \pm \frac{\|\theta\|_{\ell^1}}{n_t}$$

(by lemma C.1). Furthermore, $\mathcal{D}_k \leq \sqrt{2}|k - k_0|$ since $\psi_j = \sqrt{2}\cos(2\pi j \cdot) : [0; 1] \to [-\sqrt{2}; \sqrt{2}]$. By [2, lemma 14], with probability greater than $1 - e^{-y}$, for any $\varepsilon > 0$,

$$\left| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{\mathcal{D}_k}{n_t} \right| \leq \varepsilon \frac{\mathcal{D}_k}{n_t} + \kappa \left(\frac{\|s\|_\infty [\log n + y]}{(\varepsilon \wedge 1) n_t} + \frac{|k - k_0| [\log n + y]^2}{(\varepsilon \wedge 1)^3 n_t^2} \right).$$

Let $\varepsilon_1 = \sqrt{\frac{\|s\|_\infty (\log n + y)}{|k - k_0|}} \wedge 1$. If $\varepsilon_1 = 1$, then $|k - k_0| \leq \|s\|_\infty (y + \log n)$ therefore $\varepsilon_1 \frac{|k-k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} \leq (1 + \kappa) \frac{\|s\|_\infty (y + \log n)}{n_t}$. If $\varepsilon_1 < 1$, then

$$\varepsilon_1 \frac{|k - k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} = (1 + \kappa) \sqrt{\|s\|_\infty (y + \log n)} \frac{\sqrt{|k - k_0|}}{n_t}.$$

In all cases, if $k > k_0$,

$$\varepsilon_1 \frac{|k - k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} \leq (1 + \kappa) \|s\|_\infty (y + \log n) \frac{\sqrt{|k - k_0|}}{n_t}. \tag{61}$$

Let $\varepsilon_2 = \frac{\sqrt{\log n + y}}{n_t^{\frac{1}{4}}} \wedge 1$. If $\frac{\sqrt{y + \log n}}{n_t^{\frac{1}{4}}} \geq 1 = \varepsilon_2$, then

$$\varepsilon_2 \frac{|k - k_0|}{n_t} + \kappa \frac{|k - k_0| [\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} \leq \sqrt{y + \log n} \frac{|k - k_0|}{n_t^{\frac{5}{4}}} + \kappa \frac{|k - k_0| (y + \log n)^2}{n_t^2} \leq (1 + \kappa)(y + \log n)^2 \frac{|k - k_0|}{n_t^{\frac{5}{4}}}.$$

If $\varepsilon_2 = \frac{\sqrt{y + \log n}}{n_t^{\frac{1}{4}}} < 1$, then

$$\varepsilon_2 \frac{|k - k_0|}{n_t} + \kappa \frac{|k - k_0| [\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} = \sqrt{y + \log n} \frac{|k - k_0|}{n_t^{\frac{5}{4}}} + \kappa (y + \log n)^2 \frac{|k - k_0|}{n_t^2} \frac{n_t^{\frac{3}{4}}}{(y + \log n)^{\frac{3}{2}}}$$

$$\leq (1 + \kappa) \sqrt{y + \log n} \frac{|k - k_0|}{n_t^{\frac{5}{4}}}.$$

In all cases,

$$\varepsilon_2 \frac{|k-k_0|}{n_t} + \kappa \frac{|k-k_0|[\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} \leq (1+\kappa)(y+\log n)^2 \frac{|k-k_0|}{n_t^{\frac{5}{4}}}. \tag{62}$$

By lemma C.2,

$$\begin{aligned}
\Big| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{\mathcal{D}_k}{n_t} \Big| &\leq \inf_{\varepsilon \geq 0} \left\{ \varepsilon \frac{\mathcal{D}_k}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon \wedge 1) n_t} \right\} \\
&\quad + \inf_{\varepsilon \geq 0} \left\{ \varepsilon \frac{\mathcal{D}_k}{n_t} + \kappa \frac{|k-k_0|[\log n + y]^2}{(\varepsilon \wedge 1)^3 n_t^2} \right\} \\
&\leq \varepsilon_1 \frac{|k-k_0|}{n_t} + \kappa \frac{\|s\|_\infty [\log n + y]}{(\varepsilon_1 \wedge 1) n_t} + \varepsilon_2 \frac{|k-k_0|}{n_t} \\
&\quad + \kappa \frac{|k-k_0|[\log n + y]^2}{(\varepsilon_2 \wedge 1)^3 n_t^2} + (\varepsilon_1 + \varepsilon_2) \frac{\|\theta\|_{\ell^1}}{n_t} \\
&\leq (1+\kappa) \|s\|_\infty (y + \log n) \frac{\sqrt{|k-k_0|}}{n_t} \\
&\quad + (1+\kappa)(y+\log n)^2 \frac{|k-k_0|}{n_t^{\frac{5}{4}}} + \frac{2\|\theta\|_{\ell^1}}{n_t},
\end{aligned}$$

by equations (61), (62). In conclusion, on an event E_y of probability greater than $1 - e^{-y}$,

$$\begin{aligned}
\Big| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{|k-k_0|}{n_t} \Big| &\leq \Big| \sum_{j=k_0+1}^{k} (\hat{\theta}_j^T - \theta_j)^2 - \frac{\mathcal{D}_k}{n_t} \Big| + \frac{\|\theta\|_{\ell^1}}{n_t} \\
&\leq \frac{3\|\theta\|_{\ell^1}}{n_t} + (1+\kappa)\|s\|_\infty (y + \log n) \\
&\quad \times \left[\frac{\sqrt{|k-k_0|}}{n_t} + (y+\log n) \frac{|k-k_0|}{n_t^{\frac{5}{4}}} \right].
\end{aligned}$$

This proves equation (58).

If $k_0 = k_* + \alpha_1 \Delta$ and $k = k_* + \alpha_2 \Delta$, then by hypothesis 5 of Theorem 1,

$$\begin{aligned}
\frac{\sqrt{|k-k_0|}}{n_t} &= \sqrt{|\alpha_2 - \alpha_1|} \sqrt{\frac{\Delta}{n_v n_t}} \sqrt{\frac{n_v}{n_t}} \\
&= \sqrt{|\alpha_2 - \alpha_1|} \sqrt{\frac{n - n_t}{n_t}} \mathfrak{e} \\
&\leq \sqrt{|\alpha_2 - \alpha_1|} n^{-\frac{\delta_4}{2}} \mathfrak{e} \qquad (63) \\
&\leq \left(\sqrt{|\alpha_1|} + \sqrt{|\alpha_2|} \right) n^{-\frac{\delta_4}{2}} \mathfrak{e} \\
&\leq (\max(1, f_n(\alpha_1)) + \max(1, f_n(\alpha_2)) \, n^{-\frac{\delta_4}{2}} \mathfrak{e}. \qquad (64)
\end{aligned}$$

Moreover, by lemma A.1,

$$\begin{aligned}\frac{|k-k_0|}{n_t^{\frac{5}{4}}} &= \frac{|\alpha_2-\alpha_1|\Delta}{n_t}\frac{1}{n_t^{\frac{1}{4}}}\\ &= \frac{|\alpha_2-\alpha_1|}{n_t^{\frac{1}{4}}}\mathcal{E}\\ &\leq (\max(1,f_n(\alpha_1))+\max(1,f_n(\alpha_2)))\,\mathfrak{e}\frac{\sqrt{2n_v or(n_t)+2}}{n_t^{\frac{1}{4}}}\end{aligned}$$

Since

$$\frac{\|\theta\|_{\ell^1}}{n_t} = \frac{n_v}{n_t}\frac{\|\theta\|_{\ell^1}}{n_v} \leq \|\theta\|_{\ell^1}\frac{n-n_t}{n_t}\mathfrak{e} \leq \|\theta\|_{\ell^1} n^{-\delta_4}\mathfrak{e}, \tag{65}$$

equation (59) follows from equations (58) and (63).

Let $k_1 = \lceil n_t^{\frac{1}{3+\delta_1}}\rceil$, so that $n_t^{\frac{1}{3+\delta_1}} \leq k_1 \leq 2n_t^{\frac{1}{3+\delta_1}}$. By hypothesis 1 of Theorem 1, $\sum_{j=k+1}^{+\infty}\theta_j^2 \leq \frac{c_1}{k^{2+\delta_1}}$ therefore

$$\mathrm{or}(n_t) \leq \inf_{k\in\mathbb{N}^*}\frac{c_1}{k^{2+\delta_1}}+\frac{k}{n_t} \leq \frac{c_1}{k_1^{2+\delta_1}}+\frac{k_1}{n_t} \leq \frac{c_1}{n_t^{\frac{2}{3+\delta_1}}}+\frac{2n_t^{\frac{1}{3+\delta_1}}}{n_t} \leq \frac{2+c_1}{n_t^{\frac{2}{3+\delta_1}}}.$$

Thus $2+2n_v\mathrm{or}(n_t) \leq 2+2n_t\mathrm{or}(n_t) \leq (5+2c_1)n_t^{\frac{1+\delta_1}{3+\delta_1}}$, hence

$$\begin{aligned}\frac{|k-k_0|}{n_t^{\frac{5}{4}}} &\leq (\max(1,f_n(\alpha_1))+\max(1,f_n(\alpha_2)))\,\mathfrak{e}\sqrt{5+2c_1}\frac{n_t^{\frac{1+\delta_1}{6+2\delta_1}}}{n_t^{\frac{1}{4}}}\\ &\leq [\max(1,f_n(\alpha_1))+\max(1,f_n(\alpha_2))]\,\sqrt{5+2c_1}n_t^{-\frac{1}{12}}\mathfrak{e}\\ &\leq [\max(1,f_n(\alpha_1))+\max(1,f_n(\alpha_2))]\,\sqrt{5+2c_1}\frac{2^{\frac{1}{12}}}{n^{\frac{1}{12}}}\mathfrak{e}.\end{aligned} \tag{66}$$

Finally,. Equation (60) follows from equations (58), (64) and (66).

□

Lemme C.4. *Let $(c_{i,j})_{(i,j)\in\mathbb{N}^2}$ be real coefficients. Let $I_1, I_2 \subset \mathbb{N}$ be two finite sets. Let $(\theta_j)_{j\in\mathbb{N}}$ be a sequence. Let $C = \max\left\{\sup_{i\in I_1}\sum_{j\in I_2}|c_{i,j}|, \sup_{i\in I_2}\sum_{j\in I_1}|c_{i,j}|\right\}$. Then*

$$\sum_{i\in I_1}\left(\sum_{j\in I_2}c_{i,j}\theta_j\right)^2 \leq C^2\sum_{j\in I_2}\theta_j^2$$

and

$$\left|\sum_{i\in I_1}\sum_{j\in I_2}\theta_i\theta_j c_{i,j}\right| \leq C\sqrt{\left(\sum_{i\in I_1}\theta_i^2\right)\left(\sum_{j\in I_2}\theta_j^2\right)}.$$

Proof. Let $C_i = \sum_{j\in I_2} |c_{i,j}|$. Then

$$\begin{aligned}
\sum_{i\in I_1}\left(\sum_{j\in I_2} c_{i,j}\theta_j\right)^2 &= \sum_{i\in I_1} C_i^2\left(\frac{1}{C_i}\sum_{j\in I_2}\operatorname{sgn}(c_{i,j})|c_{i,j}|\theta_j\right)^2 \\
&\leq \sum_{i\in I_1}\frac{C_i^2}{C_i}\sum_{j\in I_2}|c_{i,j}|\theta_j^2 \text{ by the Jensen inequality} \\
&\leq \left(\max_{i\in I_1} C_i\right)\sum_{j\in I_2}\theta_j^2\sum_{i\in I_1}|c_{i,j}| \\
&\leq C^2\sum_{j\in I_2}\theta_j^2.
\end{aligned}$$

This proves the first equation. Furthermore, for any $\lambda > 0$

$$\begin{aligned}
\left|\sum_{i\in I_1}\sum_{j\in I_2}\theta_i\theta_j c_{i,j}\right| &\leq \sum_{i\in I_1}\sum_{j\in I_2}\left(\frac{\lambda}{2}\theta_i^2 + \frac{\theta_j^2}{2\lambda}\right)|c_{i,j}| \\
&= \frac{\lambda}{2}\sum_{i\in I_1}\theta_i^2\sum_{j\in I_2}|c_{i,j}| + \frac{1}{2\lambda}\sum_{j\in I_2}\theta_j^2\sum_{i\in I_1}|c_{i,j}| \\
&\leq C\left(\frac{\lambda}{2}\sum_{i\in I_1}\theta_i^2 + \frac{1}{2\lambda}\sum_{j\in I_2}\theta_j^2\right),
\end{aligned}$$

which proves the second equation by optimizing over $\lambda > 0$. □

Lemme C.5. *Under the assumptions of Theorem 1, there exists a constant $\kappa(c_1, c_2) > 0$ such that for any $x > 0$,*

$$(b_x - a_x)\Delta \leq \kappa(1+x)n^{\frac{1}{3}} \tag{67}$$

$$k_* + a_x\Delta \geq \frac{\kappa}{(1+x)^{\frac{1}{\delta_2}}}n_t^{\frac{2}{3\delta_2}} \tag{68}$$

$$k_* + b_x\Delta \leq \kappa(1+x)n^{\frac{1}{3}} \tag{69}$$

Proof. First, remark that by hypothesis 1 of Theorem 1,

$$\operatorname{or}(n_t) \leq \min_{k\in\mathbb{N}}\frac{c_1}{k^2} + \frac{k}{2n_t} \leq \frac{3c_1^{\frac{1}{3}}}{n_t^{\frac{2}{3}}}. \tag{70}$$

Remark that by definition, $-a_x\Delta \leq k_* \leq n_t or(n_t)$ and

$$\begin{aligned} x\left[2or(n_t)+\frac{1}{n-n_t}\right] &\geq \mathfrak{e}f_n(b_x) \\ &= \sum_{j=k_*+1}^{k_*+b_x\Delta}\left[\frac{1}{n_t}-\theta_j^2\right] \\ &= \frac{b_x\Delta}{n_t}-\sum_{j=k_*+1}^{k_*+b_x\Delta}\theta_j^2 \\ &\geq \frac{b_x\Delta}{n_t}-or(n_t). \end{aligned}$$

It follows that

$$b_x\Delta \leq (2x+1)n_t or(n_t)+\frac{x}{n-n_t}. \tag{71}$$

By equation (70) and hypothesis (6), there exists some constant $\kappa > 0$ such that

$$(b_x-a_x)\Delta \leq 2(x+1)n_t or(n_t)+\frac{x}{n-n_t} \leq \kappa(x+1)n^{\frac{1}{3}}.$$

This proves equation (67).

By hypothesis 2 of Theorem 1 and by definition of a_x,

$$\begin{aligned} c_2(k_*+a_x\Delta)^{-\delta_2} &\leq \sum_{j=k_*+a_x\Delta+1}^{+\infty}\theta_j^2 \\ &= \sum_{j=k_*+a_x\Delta+1}^{+\infty}\left[\theta_j^2-\frac{1}{n_t}\right]+\frac{|a_x\Delta|}{n_t}+\sum_{j=k_*+1}^{+\infty}\theta_j^2 \\ &\leq or(n_t)+\mathfrak{e}f_n(a_x) \\ &\leq (1+2x)or(n_t)+\frac{x}{n-n_t}. \end{aligned} \tag{72}$$

Since by hypothesis 6 of Theorem 1, $n_v \geq n^{\frac{2}{3}+\delta_5}$, Equation (72) yields

$$c_2(k_*+a_x\Delta)^{-\delta_2} \leq (1+2x)\mathrm{or}(n_t)+\frac{x}{n^{\frac{2}{3}}}.$$

By equation (70), it follows that, for some constant $\kappa(c_1,c_2)$,

$$k_*+a_x\Delta \geq \frac{\kappa}{(1+x)^{\frac{1}{\delta_2}}}n_t^{\frac{2}{3\delta_2}}.$$

This proves equation (68).

Equation (69) follows from equation (71) and the fact that $k_* \leq n_t or(n_t)$. □

Claim 9. *Let*

$$u_2 = \min\left(\delta_4(1+\delta_6-\delta_3), \frac{2\delta_6}{3\delta_2}\right) > 0.$$

Let $\alpha_g \geq a_x$ and $\alpha_d \leq b_x$. For some constant $\kappa \leq c_3 + c_3 c_6$ and all $j \in \{a_x\Delta + 1, \ldots, b_x\Delta, \}$

$$\theta^2_{k_*+j} \leq c_3 c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t} + \kappa(1+x)^{\frac{\delta_6}{\delta_2}} \left[\frac{|j|}{\Delta} + \frac{\mathbb{I}\{j<0\}}{\mathcal{E}} \sum_{i=k_*-(j)_-+1}^{k_*} \theta_i^2 \right] n^{-u_2}\mathcal{E}. \tag{73}$$

Proof. Let $l = \lfloor c_3(k_*+j)^{\delta_3} \rfloor$. By hypothesis 3 of Theorem 1,

$$\theta^2_{k_*+j} \leq c_6(k_*+j)^{-\delta_6} \sum_{i=j}^{j+l} \theta^2_{k_*+i}.$$

We distinguish two cases.

- If $j > 0$, then by definition of k_*,

$$\begin{aligned}
\theta_j^2 &\leq c_6(k_*+j)^{-\delta_6} \sum_{i=j}^{j+l} \theta^2_{k_*+i} \\
&\leq c_6(k_*+j)^{-\delta_6} \sum_{i=1}^{j+l} \theta^2_{k_*+i} \\
&\leq c_6(k_*+j)^{-\delta_6} \frac{j+l}{n_t} \\
&\leq c_3 c_6 \frac{(k_*+j)^{\delta_3-\delta_6}}{n_t} + c_6(k_*+j)^{-\delta_6} \frac{j}{n_t} \\
&\leq (k_*+j)^{\delta_3-\delta_6} \frac{c_3 c_6}{n_t} + c_6 k_*^{-\delta_6} \frac{j}{\Delta}\mathcal{E}.
\end{aligned}$$

  Since $\delta_3 - \delta_6 < 1$,

$$(k_*+j)^{\delta_3-\delta_6} \leq k_*^{\delta_3-\delta_6} + j^{\delta_3-\delta_6} \leq k_*^{\delta_3-\delta_6} + \max\left(1, \frac{j}{\Delta}\right) \frac{\Delta}{\Delta^{1+\delta_6-\delta_3}}$$

  By assumption (5) and lemma 5.3, $\Delta \geq n^{\delta_4}$, which yields

$$\theta^2_{k_*+j} \leq c_3 c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t} + c_3 c_6 n^{-(1+\delta_6-\delta_3)\delta_4} \max\left(1, \frac{j}{\Delta}\right) \mathcal{E} + c_6 k_*^{-\delta_6} \frac{j}{\Delta}\mathcal{E}.$$

- If $j \leq 0$, then by definition of k_*,

$$\begin{aligned}
\theta^2_{k_*+j} &\leq c_6(k_*+j)^{-\delta_6} \left(\sum_{i=j}^{0} \theta^2_{k_*+i} + \frac{(j+l)_+}{n_t} \right) \\
&\leq c_6(k_*+j)^{-\delta_6} \left(\sum_{i=j}^{0} \theta^2_{k_*+i} \right) + c_6(k_*+j)^{-\delta_6} \frac{l}{n_t} \\
&\leq c_6(k_*+j)^{-\delta_6} \left(\sum_{i=j}^{0} \theta^2_{k_*+i} \right) + c_3 c_6 \frac{(k_*+j)^{\delta_3-\delta_6}}{n_t}
\end{aligned}$$

Since $j \geq a_x$ and $j \leq 0$, this yields

$$\theta^2_{k_*+j} \leq c_6(k_* + a_x\Delta)^{-\delta_6} \left(\sum_{i=j}^{0} \theta^2_{k_*+i} \right) + c_3 c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t}.$$

By the previous lemma (lemma C.5),

$$k_* \geq k_* + a_x\Delta \geq \frac{\kappa}{(1+x)^{\frac{1}{\delta_2}}} n^{\frac{2}{3\delta_2}}.$$

Thus, combining both cases yields

$$\theta^2_{k_*+j} \leq c_3 c_6 \frac{k_*^{\delta_3-\delta_6}}{n_t} + \kappa(1+x)^{\frac{\delta_6}{\delta_2}} n^{-u_2} \left[\frac{(j)_-}{\Delta}\mathcal{E} + \sum_{i=k_*-(j)_-}^{k_*} \theta_i^2 \right]$$

for all $j \in \{k_* + a_x\Delta + 1, \ldots, k_* + b_x\Delta\}$, where $\kappa \leq c_3 + c_3c_6$ is a constant and

$$u_2 = \min\left(\delta_4(1 + \delta_6 - \delta_3), \frac{2\delta_6}{3\delta_2} \right) > 0.$$

This proves claim 9. □

Acknowledgements

The author would like to thank Sylvain Arlot, whose pertinent advice led to improvements in the presentation of the paper.

Funding

While writing this article, the author received funding from the European Union's Horizon 2020 research program under grant agreement N° 811017.

References

[1] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010.

[2] Sylvain Arlot and Matthieu Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research (JMLR)*, 17(208):1–50, 2016.

[3] Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *arXiv e-prints*, page arXiv:2001.11111, January 2020.

[4] Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16339–16350. Curran Associates, Inc., 2020.

[5] Philippe Berthet and David M. Mason. Revisiting two strong approximation results of dudley and philipp. In *High Dimensional Probability*, pages 155–172. Institute of Mathematical Statistics, 2006.

[6] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

[7] Alain Célisse and Stephane Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2250–2368, 2008.

[8] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564 – 1597, 2014.

[9] Miklós Csörgő and Lajos Horváth. A note on strong approximations of multivariate empirical processes. *Stochastic Processes and their Applications*, 28(1):101–109, 1988.

[10] Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[11] Sam Efromovich. *Nonparametric curve estimation*. Springer Series in Statistics. Springer, New York, NY, mar 1999.

[12] Peter Hall. Laws of the iterated logarithm for nonparametric density estimators. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 56(1):47–61, Mar 1981.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, 2009.

[14] Vladimir I. Koltchinskii. Komlos-major-tusnady approximation for the general empirical process and haar expansions of classes of functions. *Journal of Theoretical Probability*, 7(1):73–118, Jan 1994.

[15] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32(1):111–131, Mar 1975.

[16] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. *The Annals of Statistics*, 39, 11 2009.

[17] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55, 2021.

[18] Fabien Navarro and Adrien Saumard. Slope heuristics and v-fold model selection in heteroscedastic regression using strongly localized bases. *ESAIM: Probability and Statistics*, 21:412–451, 2017.

[19] Emmanuel Rio. Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98(1):21–45, Mar 1994.

[20] Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3):351–371, 2006.